

Hybrid Pathwise Sensitivity Methods for Discrete Stochastic Models of Chemical Reaction Systems

Elizabeth Skubak Wolf*, David F. Anderson†

November 19, 2014

Abstract

Stochastic models are often used to help understand the behavior of intracellular biochemical processes. The most common such models are continuous time Markov chains (CTMCs). Parametric sensitivities, which are derivatives of expectations of model output quantities with respect to model parameters, are useful in this setting for a variety of applications. In this paper, we introduce a class of hybrid pathwise differentiation methods for the numerical estimation of parametric sensitivities. The new hybrid methods combine elements from the three main classes of procedures for sensitivity estimation, and have a number of desirable qualities. First, the new methods are unbiased for a broad class of problems. Second, the methods are applicable to nearly any physically relevant biochemical CTMC model. Third, and as we demonstrate on several numerical examples, the new methods are quite efficient, particularly if one wishes to estimate the full gradient of parametric sensitivities. The methods are rather intuitive and utilize the multilevel Monte Carlo philosophy of splitting an expectation into separate parts and handling each in an efficient manner.

1 Introduction

New methods for the estimation of parametric sensitivities are introduced that are applicable to a class of stochastic models widely utilized in the biosciences. In particular, the theoretical analysis and algorithms provided here extend the validity of the pathwise method developed by Sheppard, Rathinam, and Khammash [30], with related earlier work by Glasserman [15], to nearly all physically relevant stochastic models from biochemistry. The extension is achieved by providing an explicit, numerically computable term for the bias introduced by standard pathwise differentiation methods. The methods developed here are naturally unbiased and are relatively easy to implement. Furthermore, they are quite efficient, in some cases providing a speed up of multiple orders of magnitude over the previous state of the art.

1.1 Mathematical model and problem statement

Mathematical model. We consider the parametrized family of continuous time Markov chain (CTMC) models satisfying the stochastic equation

$$X_\theta(t) = X_\theta(0) + \sum_{k=1}^K Y_k \left(\int_0^t \lambda_k(\theta, X_\theta(s)) ds \right) \zeta_k, \quad (1)$$

where the state space \mathcal{S} of X_θ is a subset of \mathbb{Z}^d , $K < \infty$, the $\{Y_k\}$ are independent unit-rate Poisson processes, $\theta \in \mathbb{R}^R$ is a vector of model parameters, and where for each $k \in \{1, \dots, K\}$ we have a fixed reaction vector $\zeta_k \in \mathbb{Z}^d$ and a nonnegative intensity, or propensity, function $\lambda_k : \mathbb{R}^R \times \mathbb{Z}^d \rightarrow \mathbb{R}_{\geq 0}$. Such models are used extensively in the study of biochemical processes [6, 7, 10, 14, 21, 25, 27, 32] in which case the vectors ζ_k can be decomposed into the difference between the *source vector* $\nu_k \in \mathbb{Z}_{\geq 0}^d$, giving the numbers of molecules

*Saint Mary's College, ewolf@saintmarys.edu.

†University of Wisconsin at Madison, anderson@math.wisc.edu.

required for a given reaction to proceed, and the *product vector* $\nu'_k \in \mathbb{Z}_{\geq 0}^d$, giving the numbers of molecules produced by a given reaction. Specifically, in this case $\zeta_k = \nu'_k - \nu_k$. Under the assumption of mass action kinetics, which assumes intensities of the form

$$\lambda_k(\theta, x) = \theta_k \prod_{i=1}^d \frac{x_i!}{(x_i - \nu_{ki})!} 1_{\{x - \nu_k \geq 0\}}, \quad \text{for } x \in \mathbb{Z}_{\geq 0}^d, \quad (2)$$

the parameter vector θ commonly represents some subset of the rate constants $\{\theta_k\}$ of the K reactions. Note that in the biochemical setting the state space \mathcal{S} is a subset of $\mathbb{Z}_{\geq 0}^d$.

Models of the form (1) satisfy the Kolmogorov forward equation, which is typically called the chemical master equation in the biology and chemistry literature,

$$\frac{d}{dt} p_\pi^\theta(t, x) = \sum_{k=1}^K p_\pi^\theta(t, x - \zeta_k) \lambda_k(\theta, x - \zeta_k) 1_{\{x - \zeta_k \in \mathcal{S}\}} - \sum_{k=1}^K p_\pi^\theta(t, x) \lambda_k(\theta, x), \quad (3)$$

where $p_\pi^\theta(t, x)$ is the probability the state of the system is $x \in \mathcal{S}$ at time $t \geq 0$ given an initial distribution of π . The infinitesimal generator for the CTMC (1) is the operator \mathcal{A}^θ defined via

$$(\mathcal{A}^\theta f)(x) = \sum_{k=1}^K \lambda_k(\theta, x) (f(x + \zeta_k) - f(x)), \quad (4)$$

for $f : \mathbb{R}^d \rightarrow \mathbb{R}$ vanishing off a finite set [11]. For more background on this model see [6, 7, 21, 22].

We note that many lattice-valued processes can be represented similarly to (1), where a counting process is used to determine the number of jumps that have taken place in one of finitely many specified directions. In particular, models satisfying (1) also arise in queueing theory and the study of population processes. As biochemical reaction networks are the main motivation for this work, we use biochemical terminology and examples throughout, and simply note that the presented methods are also applicable in those other settings.

Problem statement. The process X_θ satisfying (1) is right continuous and has left hand limits. That is, X_θ is càdlàg and is an element of the Skorohod space $D_{\mathbb{Z}^d}[0, \infty)$. Consider the output quantity of the CTMC model (1) given by $\mathbb{E}[f(\theta, X_\theta)]$, where $f : \mathbb{R}^R \times D_{\mathbb{Z}^d}[0, \infty) \rightarrow \mathbb{R}$ is some measurable function of θ and X_θ . We are interested in the problem of numerically computing the gradient $\nabla_\theta \mathbb{E}[f(\theta, X_\theta)]$ for a wide class of functionals f . Specifically, we are interested in functionals of the form

$$f(\theta, X_\theta) = h(\theta, X_\theta(t)), \quad \text{for } t \text{ fixed}, \quad (5)$$

where $h : \mathbb{R}^R \times \mathbb{Z}^d \rightarrow \mathbb{R}$, or path functionals of the form

$$L(\theta) := \int_a^b F(\theta, X_\theta(s)) ds, \quad (6)$$

where $0 \leq a \leq b < \infty$ and $F : \mathbb{R}^R \times \mathbb{Z}^d \rightarrow \mathbb{R}$. We will write $L_X(\theta)$ for $L(\theta)$ when we wish to be clear about the underlying process X , and will denote $J(\theta) := \mathbb{E}[L(\theta)]$.

We will focus most of our attention on functionals of the form (6) as we will show in Section 2.2.1 how basic smoothing procedures allow us to use such functionals in conjunction with our new methods to provide estimates for $\nabla_\theta \mathbb{E}[f(\theta, X_\theta)]$ when f is of the form (5). Thus, under some mild regularity conditions on the functions λ_k and F (see Conditions 1 and 2 in this section below), we focus on the problem of estimating

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}[L(\theta)] = \left[\frac{\partial}{\partial \theta_i} \mathbb{E} \left(\int_a^b F(\theta, X_\theta(s)) ds \right) \right]_{i=1, \dots, R} \quad (7)$$

at some fixed value $\theta_0 \in \mathbb{R}^R$. We will generally write θ rather than θ_0 if the context is clear.

1.2 A brief review of methods

Due to the importance of having reliable numerical estimators for gradients, there has recently been a plethora of research articles focusing on their development and analysis [2, 5, 18, 20, 24, 26, 28, 30, 31]. There are three main classes of methods that carry out the task of estimating these derivatives: finite difference methods, likelihood ratio methods, and pathwise methods. Each class has its own benefits and drawbacks.

- Estimators built via **finite differences** are easy to implement and often have a low variance. However, these estimators provide a biased estimate [2, 28, 31]. See Section 2.1.
- Estimators built using **likelihood ratios** are unbiased, but often have a high variance [2, 26]. The use of the usual weight function as a control variate can lower the variance, sometimes dramatically so. See Section 2.3.
- **Pathwise methods**, known as (infinitesimal) perturbation analysis in the discrete event systems literature [15, 17], are unbiased and are often quite fast [30]. Unfortunately, biochemical models only rarely satisfy the conditions required for the applicability of these methods. See, for example, the appendix of [30] and Section 2.2 below. Greatly expanding the applicability of the pathwise methods already developed for biochemical processes, for example in [30], is one of the main contributions of this work.

In some recent works Gupta and Khammash have developed a new type of method that does not fit neatly into one of the above categories [18, 19]. Their new method, the Poisson path approximation (PPA) method, which is an improvement on their auxiliary path approximation (APA) method introduced in [18], is unbiased and is quite efficient [19]. This method does, however, require additional simulation of partial paths, which may significantly reduce efficiency on some models.

1.3 A high level overview of the present work

Elements from each of the three general classes of methods outlined in Section 1.2 above will be utilized in the development of estimators that combine the strengths of each. Further, the methods introduced here utilize the multilevel Monte Carlo philosophy by splitting a desired quantity into pieces, and then handling each piece separately and efficiently [4, 13]. Specifically, much of the computational work is carried out with a pathwise method [30] applied to an approximate process, ensuring the overall method is efficient. In order to correct for the bias introduced by the use of an approximate process, the gradient of an error term is computed. The error term is represented as the expectation of a function of a coupling between the original process and the approximate process. The likelihood ratio method is used to compute the necessary derivative on this error term. The coupling used between the exact and approximate processes is the split coupling [2, 5].

Expanding upon the previous paragraph, we give a high level summary of the new method as applied to the functional $L_X(\theta)$ in (6). First note that by adding and subtracting the appropriate terms,

$$\mathbb{E} \left[\int_a^b F(\theta, X_\theta(s)) ds \right] = \mathbb{E} \left[\int_a^b (F(\theta, X_\theta(s)) - F(\theta, Z_\theta(s))) ds \right] + \mathbb{E} \left[\int_a^b F(\theta, Z_\theta(s)) ds \right],$$

where Z_θ is any process that can be built on the same probability space as X_θ , and where we assume the expectations above are finite. Then, assuming the derivatives exist,

$$\nabla_\theta \mathbb{E} \left[\int_a^b F(\theta, X_\theta(s)) ds \right] = \nabla_\theta \mathbb{E} \left[\int_a^b (F(\theta, X_\theta(s)) - F(\theta, Z_\theta(s))) ds \right] + \nabla_\theta \mathbb{E} \left[\int_a^b F(\theta, Z_\theta(s)) ds \right]. \quad (8)$$

We are now able to use different methods to compute the two derivatives on the right-hand side of the above equation. We have complete control over Z_θ , and we will construct it so that (i) pathwise methods may be utilized for the final derivative on the right-hand side of (8), and (ii) Z_θ is a good approximation to X_θ . The error term, which is the first term on the right-hand side of (8), will be estimated via a likelihood ratio method. The efficiency of the overall method rests upon two facts. First, the error term can be quickly

estimated because its variance will be small if Z_θ is a good approximation to X_θ . This helps overcome the often problematically large variance of a likelihood estimator. Second, the final derivative can be estimated quickly because pathwise methods are fast when they are applicable.

In this paper we present what we believe is a reasonable choice for the process Z_θ in (8). Specifically, it will have the same jump directions $\{\zeta_k\}$ as X_θ , but different intensity functions and an enlarged state space. While we hope to impart why we believe it to be a good choice, many other options for Z_θ exist and can be explored in future research. Improvements in the selection of the process Z_θ will correspond with improvements to the overall method. The use of multilevel Monte Carlo on either of the needed derivatives could also lead to a significant improvement in efficiency.

Our numerical examples section shows that the methods we introduce here fit well into the group of existing methods for the numerical estimation of parametric sensitivities in the jump process setting. They are quite efficient on all examples, sometimes significantly more efficient than any other existing method. However, and not surprisingly given the amount of effort that has been put into development over the past few years, they are not *always* the most efficient. In particular, sometimes PPA (Gupta and Khammash, [19]) or the coupled finite difference method (Anderson, [2]) is most efficient. With such a strong group of methods having been developed over the past few years, we feel future work in the field should also include the determination of which methods to use for different model and problem types.

1.4 Regularity conditions

We end this introduction with two regularity conditions which are necessary for the validity of the methods introduced here. The first condition guarantees that solutions to equation (1) exist for all time. The second condition relates to F of (6), and simply ensures that F does not grow too fast in the x variable. Both conditions are required in our proofs in Appendix A. Conditions to be satisfied by the approximate process Z_θ will be developed as needed throughout the paper. In particular, see Conditions 3, 4, and 5.

For $x \in \mathbb{Z}^d$ we use the notation $\|x\|$ to denote the 1-norm, $\|x\| = \sum_{i=1}^d |x_i|$.

Definition. We say that $h : \mathbb{R}^R \times \mathcal{S} \rightarrow \mathbb{R}$ has **uniform polynomial growth at θ** if there is a neighborhood $\Theta \subset \mathbb{R}^R$ of θ and constants $C, p > 0$ such that $|\sup_{\theta \in \Theta} h(\theta, x)| \leq C(1 + \|x\|^p)$ for all $x \in \mathcal{S}$. If p may be taken to be 1, we say that h has **uniform linear growth at θ** .

Let $\mathbf{1}$ denote the vector of all ones. Define $\mathcal{R}_1 \subset \{1, \dots, K\}$ so that $k \in \mathcal{R}_1$ if and only if $\mathbf{1} \cdot \zeta_k > 0$. Define $\mathcal{R}_2 = \{1, \dots, K\} \setminus \mathcal{R}_1$. Note that if $\mathcal{S} \subset \mathbb{Z}_{\geq 0}^d$, then \mathcal{R}_1 contains the indices of those reactions that increase the total population, i.e.

$$\|x + \zeta_k\| > \|x\|, \quad \text{for all } x \in \mathcal{S},$$

while reactions with indices in \mathcal{R}_2 either decrease the total population or leave it unchanged.

Condition 1. The intensities λ_k satisfy this condition at θ if there is some neighborhood $\Theta \subset \mathbb{R}^R$ of θ such that:

1. for each $k \in \{1, \dots, K\}$ and $\theta \in \Theta$, the function λ_k has uniform polynomial growth at θ ;
2. for each $k \in \{1, \dots, K\}$, $i \in \{1, \dots, R\}$, and $\theta \in \Theta$, the function $\frac{\partial}{\partial \theta_i} \lambda_k$ exists and has uniform polynomial growth at θ ;
3. for each $k \in \mathcal{R}_1$ and $\theta \in \Theta$, the function λ_k has uniform linear growth at θ ;
4. there exist constants p and C such that for all $k \in \{1, \dots, R\}$ and all $x \in \mathcal{S}$

$$\sup_{\theta \in \Theta} \lambda_k(\theta, x) \neq 0 \Rightarrow \sup_{\theta \in \Theta} \frac{1}{\lambda_k(\theta, x)} \leq C(1 + \|x\|^p);$$

that is, for a fixed x , if the rates $\lambda_k(\theta, x)$ are not identically zero on Θ , then they must be bounded away from zero.

Note that the third part of Condition 1, which requires certain intensities to grow at most linearly, only applies to those intensity functions that are associated with transitions that increase the total population count of the system. Essentially, this portion of Condition 1 ensures that the population does not explode in finite time, and could almost certainly be weakened. We note that this condition was also utilized in [18]. Condition 1 is satisfied for most biochemical systems considered in the literature. In particular, it is satisfied by any binary chemical system with mass action kinetics.¹ For example, assuming mass action kinetics, the reactions $A \rightarrow 2A$ and $2A \rightarrow B + C$ are permissible under Condition 1. On the other hand, Condition 1 excludes $2A \rightarrow 3A$, which increases the population at a quadratic rate, and can lead to explosions.

We turn to the regularity conditions for F of (6). The following condition will allow us to bound moments of L using the moments of the process X_θ .

Condition 2. *Let $\Theta \subset \mathbb{R}^R$. The function $F : \Theta \times \mathcal{S} \rightarrow \mathbb{R}$ satisfies this condition if it is measurable, and differentiable in θ on Θ so that:*

1. *there exist constants $C_A > 1$ and $p_A > 1$ such that $\sup_{\theta \in \Theta} |F(\theta, x)| \leq C_A(1 + \|x\|^{p_A})$ for all $x \in \mathcal{S}$;*
2. *there exist constants $C_B > 1$ and $p_B > 1$ such that for all $i \in \{1, \dots, R\}$ and $x \in \mathcal{S}$ we have*

$$\sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_i} F(\theta, x) \right| \leq C_B(1 + \|x\|^{p_B}).$$

The outline for the remainder of the paper is as follows. In Section 2, we introduce the three main classes of methods for the numerical estimation of parametric sensitivities. In particular, in Section 2.2 we present Theorem 1, which gives conditions for the validity of pathwise methods for functionals of the form (6). In Section 3, we introduce an approximate process Z_θ to be utilized in (8) and formally present the new methods. In Section 4, we demonstrate several numerical results, and we present conclusions in Section 5. The proof of Theorem 1 is included in the appendix.

2 Classes of Methods

We introduce the three main classes of methods for the numerical estimation of parametric sensitivities: finite differences, pathwise derivatives, and likelihood ratios. Because the methods introduced here involve both pathwise derivatives and likelihood ratios, we discuss both in detail in Sections 2.2 and 2.3 below. Throughout these sections, we also introduce and motivate the regularity conditions and theoretical results that are required for the approximate process Z_θ of (8). In Section 3, we will combine these pieces to succinctly introduce our new method. Our main theoretical results pertaining to pathwise methods are stated in Section 2.2.4 and proven in the appendix.

2.1 Finite differences

Let $e_i \in \mathbb{R}^R$ be the vector of all zeros except a one in the i th component. Finite difference methods proceed by simply noting that for $f : \mathbb{R}^R \times D_{\mathbb{Z}^d}[0, \infty) \rightarrow \mathbb{R}$,

$$\begin{aligned} \frac{\partial}{\partial \theta_i} \mathbb{E}[f(\theta, X_\theta)] &\approx h^{-1} (\mathbb{E}[f(\theta + he_i, X_{\theta+he_i})] - \mathbb{E}[f(\theta, X_\theta)]) \\ &= h^{-1} \mathbb{E}[f(\theta + he_i, X_{\theta+he_i}) - f(\theta, X_\theta)], \end{aligned}$$

as long as the derivatives and expectations exist, and where the final equality implies the two processes have been built on the same probability space, or *coupled*. The coupling is used in order to reduce the variance of the difference between the two random variables. The two most useful couplings in the present context are the common reaction path method [28] and the split coupling method [2], the latter of which we detail explicitly in Section 2.3 in and around (24).

¹A chemical system is binary if $\sum_{i=1}^d |\nu_{ki}| \leq 2$ and $\sum_{i=1}^d |\nu'_{ki}| \leq 2$ for each $k \in \{1, \dots, K\}$.

2.2 Pathwise methods

When using a pathwise method, one begins with a probability space that does not depend on θ ; instead, one uses θ to construct the path from the underlying randomness. For our purposes, we take a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, Q)$ under which $\{Y_k, k = 1, \dots, K\}$ are independent unit-rate Poisson processes. The path X_θ is then constructed by a jump by jump procedure implied by (1), which is equivalent to an implementation of the next reaction method [1, 12]. For ease of exposition, we restrict ourselves to consideration of one element of the gradient, $\frac{\partial}{\partial \theta_i} J(\theta)$, though calculation of the full gradient can be carried out in the obvious manner.

Consider a general functional f . If the following equality holds,

$$\frac{\partial}{\partial \theta_i} \mathbb{E}[f(\theta, X_\theta)] = \mathbb{E} \left[\frac{\partial}{\partial \theta_i} f(\theta, X_\theta) \right], \quad (9)$$

then $\frac{\partial}{\partial \theta_i} \mathbb{E}[f(\theta, X_\theta)]$ can be estimated via Monte Carlo by repeated sampling of independent copies of the random variable $\frac{\partial}{\partial \theta_i} f(\theta, X_\theta)$. Unfortunately, for a wide variety of models of the form (1) and functionals f , equality in (9) does not hold. There are typically two reasons for this.

1. In many cases the random variable $\frac{\partial}{\partial \theta_i} f(\theta, X_\theta)$ is almost surely zero, in which case the right hand side of (9) is zero whereas the left hand side is not.
2. The underlying process X_θ can undergo an *interruption*, in which case $\mathbb{E} \left[\frac{\partial}{\partial \theta_i} f(\theta, X_\theta) \right]$ is typically non-zero, but still not equal to $\frac{\partial}{\partial \theta_i} \mathbb{E}[f(\theta, X_\theta)]$.

The first problem stated above commonly arises when f is a function solely of the process at the terminal time T , i.e. when $f(\theta, X_\theta) = h(X_\theta(T))$ for some $T > 0$ and $h : \mathcal{S} \rightarrow \mathbb{R}$ (as in (5) above). Then, since X_θ is a CTMC and has piecewise constant paths, $\frac{\partial}{\partial \theta_i} h(X_\theta(T)) = 0$ almost surely. This type of problem is easily overcome by any number of smoothing procedures, with a few outlined below in Section 2.2.1. The second problem, in which there is an interruption, is a more serious problem with the method. Interruptions are discussed in more detail in Section 2.2.2 below. Overcoming this type of problem while still utilizing the pathwise framework can be viewed as a major contribution of this work.

2.2.1 Smoothing

As will be seen in Section 2.2.3, pathwise methods are capable of providing estimates of derivatives of functionals of the form $\int_a^b F(\theta, X_\theta(s)) ds$, where $a, b \in \mathbb{R}$ and $F : \mathbb{R}^R \times \mathbb{Z}^d \rightarrow \mathbb{R}$ satisfies mild regularity conditions. Thus, in order to estimate derivatives of, for example, $\mathbb{E}[f(X_\theta(T))]$, where $f : \mathcal{S} \rightarrow \mathbb{R}$, one simply needs to replace $f(X_\theta(T))$ with an appropriate integral. There are a number of natural choices, with only a few discussed here.

The Regularized Pathwise Derivative (RPD) method presented in [30] estimates $\frac{\partial}{\partial \theta_i} \mathbb{E}[f(X_\theta(T))]$ using independent copies of θ -derivatives of

$$L_1(\theta) := \frac{1}{2w} \int_{T-w}^{T+w} f(X_\theta(s)) ds \approx f(X_\theta(T)), \quad (10)$$

where w is some fixed window size. Note that even when pathwise methods can be applied to the model, i.e. when there are no interruptions, this method gives a biased estimate, with the size of the bias a function of the size of w . Specifically, a smaller w leads to a smaller bias but a larger variance.

Alternatively, one may martingale methods to derive an unbiased estimator. Specifically, for a large set of functions $f : \mathbb{Z}^d \rightarrow \mathbb{R}$,

$$f(X_\theta(t)) = f(X_\theta(0)) + \int_0^t (\mathcal{A}^\theta f)(X_\theta(s)) ds + M_t^\theta, \quad (11)$$

where M_t^θ is a local martingale and \mathcal{A}^θ is the generator (4) [6, 11]. In many cases of interest M_t^θ is a martingale, in which case (4) implies

$$\mathbb{E}[f(X_\theta(t))] = \mathbb{E} \left[f(X_\theta(0)) + \int_0^t \sum_k \lambda_k(\theta, X_\theta(s)) [f(X_\theta(s) + \zeta_k) - f(X_\theta(s))] ds \right]. \quad (12)$$

For example, for processes X_θ that satisfy Condition 1, which is nearly all biologically relevant processes, equation (12) holds for functions f that grow at most polynomially. Therefore, another option for a smoothing functional would be to take

$$L_2(\theta) := f(X_\theta(0)) + \int_0^T (\mathcal{A}^\theta f)(X_\theta(s)) ds \quad (13)$$

in which case $\frac{\partial}{\partial \theta_i} \mathbb{E}[f(X_\theta(T))] = \mathbb{E}[\frac{\partial}{\partial \theta_i} L_2(\theta)]$ (see also [15], p. 176). While unbiased when it applies, this estimator tends to have higher variance than the RPD estimator so long as the parameter w is not taken too small. We shall refer to the smoothing procedure (13) as the Generator Smoothing (GS) method and will refer to the estimation procedure

$$\frac{\partial}{\partial \theta_i} \mathbb{E}[f(X_\theta(T))] \approx \frac{1}{n} \sum_{j=1}^n \frac{\partial}{\partial \theta_i} L_2^{[j]}(\theta),$$

where $L_2^{[j]}(\theta)$ is the j th independent realization of the random variable $\frac{\partial}{\partial \theta_i} L_2(\theta)$, as as the GS Pathwise method. An algorithm for the generation of the random variable $\frac{\partial}{\partial \theta_i} L_2(\theta)$ is given in Section 2.2.3 below.

2.2.2 The non-interruptive condition

Smoothing alone does not always ensure the validity of a pathwise method: for $L(\theta)$ given by (6) we still may have $\frac{\partial}{\partial \theta_i} \mathbb{E}[L(\theta)] \neq \mathbb{E}[\frac{\partial}{\partial \theta_i} L(\theta)]$. Again letting $e_i \in \mathbb{R}^R$ be the vector of all zeros except a one in the i th component, for X_θ satisfying Condition 1 it is straightforward to show that

$$\lim_{h \rightarrow 0} \mathbb{E} \left[\frac{L(\theta + h e_i) - L(\theta)}{h} \right] = \frac{\partial}{\partial \theta_i} \mathbb{E}[L(\theta)] \quad \text{and} \quad \frac{L(\theta + h e_i) - L(\theta)}{h} \xrightarrow{a.s.} \frac{\partial}{\partial \theta_i} L(\theta). \quad (14)$$

However, to have the equality

$$\frac{\partial}{\partial \theta_i} \mathbb{E}[L(\theta)] = \mathbb{E} \left[\frac{\partial}{\partial \theta_i} L(\theta) \right], \quad (15)$$

we must have convergence in mean in addition to the a.s. convergence in (14). The following condition will play a central role in achieving the convergence in mean. A similar condition was first introduced by Glasserman in the discrete event simulation literature [15]. Recall that \mathcal{S} is the state space of our process.

Condition 3 (Non-Interruptive). *The functions $\lambda_k : \Theta \times \mathcal{S} \rightarrow \mathbb{R}_{\geq 0}$, for $k \in \{1, \dots, K\}$, satisfy this condition if for each $k, \ell \in \{1, \dots, K\}$, $x \in \mathcal{S}$, and $\theta \in \Theta$, the following holds: if $\lambda_k(\theta, x) > 0$ and $\lambda_\ell(\theta, x) > 0$ for $\ell \neq k$, then $\lambda_\ell(\theta, x + \zeta_k) > 0$.*

In accordance with terminology from the discrete event simulation literature, we define an *interruption* as a change in state, from x to $x + \zeta_k$ for some k , such that for some $\ell \neq k$ we have $\lambda_\ell(\theta, x) > 0$ and $\lambda_\ell(\theta, x + \zeta_k) = 0$. If an interruption occurs, the function $L(\theta)$ can have a jump discontinuity in θ for a given realization of the process, and (15) can fail to hold. The non-interruptive Condition 3, therefore, ensures that interruptions cannot occur.

Many biological models do not satisfy Condition 3. For a simple example of a model that does not satisfy the non-interruption condition, consider the reaction network



which has reaction vectors

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Endow the system with mass action kinetics and an initial condition of precisely one A particle and zero B particles. Then the occurrence of either reaction will necessarily cause an interruption.

For models in which interruptions are possible, which includes most biochemical models, both the GS pathwise method and the RPD method may produce significant bias when estimating gradients. See Appendix B of [30] for a comment on this issue, and see Section 4 below where the bias is demonstrated numerically.

2.2.3 An algorithm for calculating $\frac{\partial}{\partial \theta_i} L(\theta)$

Providing realizations of the random variable $\frac{\partial}{\partial \theta_i} L(\theta)$, where L is of the form (6), is central to the methods presented here. This section provides the necessary numerical algorithm. The derivations are based on simulating the random time change representation (1) using the next reaction method. Conditions on the intensity functions guaranteeing that $\frac{\partial}{\partial \theta_i} \mathbb{E}[L(\theta)] = \mathbb{E}[\frac{\partial}{\partial \theta_i} L(\theta)]$ are provided in Section 2.2.4 below.

We note that the algorithm derived within this section is essentially the same as those derived in [15] and [30]. This section is included for completeness, but can be safely skipped by those familiar with pathwise differentiation.

Recalling the discussion in and around (8), the methods introduced in this article use pathwise differentiation on functionals of a non-interruptive process. This process is typically an *approximation* of the original process. Thus, in this section we denote our nominal process by Z_θ as opposed to X_θ . Further, for notational convenience in this section we take θ to be 1-dimensional.

Continuing, we suppose Z_θ is a process satisfying the stochastic equation (1) with $\theta \in \mathbb{R}$. Let $\hat{Z}_\ell(\theta)$ denote the ℓ^{th} state in the embedded discrete time chain of the process Z_θ , and let T_ℓ^θ be the ℓ^{th} jump time, with $T_0^\theta = 0$. We are interested in computing the θ -derivative of

$$L_Z(\theta) := \int_a^b F(\theta, Z_\theta(s)) ds = \sum_{\ell=0}^{N(\theta, b)} F(\theta, \hat{Z}_\ell(\theta)) [T_{\ell+1}^\theta \wedge b - T_\ell^\theta \vee a]^+, \quad (16)$$

where $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$, and where $N(\theta, b) = N$ is the number of jumps of the process through time b . If Z_θ is a non-explosive process, then $N < \infty$ with a probability of one.

The embedded chain is discrete-valued. Thus, $\frac{\partial}{\partial \theta} \hat{Z}_\ell(\theta) = 0$ a.s. wherever the derivative exists. Therefore, by (16),

$$\frac{\partial}{\partial \theta} L_Z(\theta) = \sum_{\ell=0}^N \left[[T_{\ell+1}^\theta \wedge b - T_\ell^\theta \vee a]^+ \left(\frac{\partial}{\partial \theta} F(\theta, \hat{Z}_\ell(\theta)) \right) + F(\theta, \hat{Z}_\ell(\theta)) \frac{\partial}{\partial \theta} [T_{\ell+1}^\theta \wedge b - T_\ell^\theta \vee a]^+ \right], \quad (17)$$

where the partial of the function F is with respect to the first variable. The terms involving the derivatives $\frac{\partial}{\partial \theta} F(\theta, \hat{Z}_\ell(\theta))$ are straightforward to compute. The remaining terms require the derivatives of the jump times T_ℓ^θ , so we now focus on their derivation.

Define $\Delta_\ell^\theta = T_{\ell+1}^\theta - T_\ell^\theta$ to be the holding time of the process in the ℓ^{th} state (so that the indexing begins at 0). Let $S_k^\theta(t) = \int_0^t \lambda_k(\theta, Z_\theta(s)) ds$. Note that $S_k^\theta(t)$ is the argument of the Poisson process Y_k in the stochastic equation (1). The quantity $S_k^\theta(t)$ is therefore usually referred to as the ‘internal time’ of Y_k . Let

$$I_k(t) = \inf \{ r \geq S_k^\theta(t) : Y_k(r) > Y_k(S_k^\theta(t)) \}$$

be the internal time of the first occurrence of Y_k after time $S_k(t)$. Then the holding time of the process Z_θ in the ℓ^{th} state is given by

$$\Delta_\ell^\theta = \min_k \left\{ \frac{I_k(T_\ell^\theta) - S_k^\theta(T_\ell^\theta)}{\lambda_k(\theta, \hat{Z}_\ell(\theta))} \right\}. \quad (18)$$

Let k_ℓ be the argmin in the above expression; k_ℓ is the index of the reaction that changes the system from the ℓ^{th} to the $(\ell+1)^{\text{st}}$ state. Via the product rule, we have

$$\begin{aligned} \frac{\partial}{\partial \theta} \Delta_\ell^\theta &= - \frac{I_{k_\ell} - S_{k_\ell}^\theta(T_\ell^\theta)}{\lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta))^2} \frac{\partial}{\partial \theta} \lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta)) - \lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta))^{-1} \frac{\partial}{\partial \theta} S_{k_\ell}^\theta(T_\ell^\theta) \\ &= - \frac{\Delta_\ell^\theta}{\lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta))} \frac{\partial}{\partial \theta} \lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta)) - \lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta))^{-1} \frac{\partial}{\partial \theta} S_{k_\ell}^\theta(T_\ell^\theta), \end{aligned} \quad (19)$$

where the second equality follows from (18). Note that for $t \in [T_\ell^\theta, T_{\ell+1}^\theta]$ and any $k \in \{1, \dots, K\}$ we have that $S_k^\theta(t) = S_k^\theta(T_\ell^\theta) + \lambda_k(\theta, \hat{Z}_\ell(\theta))(t - T_\ell^\theta)$. Thus

$$\frac{\partial}{\partial \theta} S_k^\theta(T_\ell^\theta) = \frac{\partial}{\partial \theta} S_k^\theta(T_{\ell-1}^\theta) + \Delta_{\ell-1}^\theta \frac{\partial}{\partial \theta} \lambda_k(\theta, \hat{Z}_{\ell-1}(\theta)) + \lambda_k(\theta, \hat{Z}_{\ell-1}(\theta)) \frac{\partial}{\partial \theta} \Delta_{\ell-1}^\theta. \quad (20)$$

The values $\{\frac{\partial}{\partial\theta}\Delta_\ell^\theta\}$ and $\{\frac{\partial}{\partial\theta}S_\ell^\theta(T_\ell^\theta)\}$ can now be solved for recursively given that $S_k^\theta(T_0^\theta) = 0$ for all k .

To find the derivatives of the T_ℓ^θ as in (17), first note that $\frac{\partial}{\partial\theta}T_0^\theta = 0$, and that for $\ell > 0$ the definition of Δ_ℓ^θ implies that

$$\frac{\partial}{\partial\theta}T_\ell^\theta = \sum_{j=0}^{\ell-1} \frac{\partial}{\partial\theta}\Delta_j^\theta.$$

Let $\ell_a \in \mathbb{N}$ be maximal such that $T_{\ell_a}^\theta \leq a$; that is, the ℓ_a^{th} jump is the last jump to occur before time a . We may now conclude that

$$\frac{\partial}{\partial\theta}[T_{\ell+1}^\theta \wedge b - T_\ell^\theta \vee a]^+ = \begin{cases} 0 & \ell < \ell_a \quad \text{or} \quad \ell > N \\ \frac{\partial}{\partial\theta}T_{\ell_a+1}^\theta = \sum_{j=0}^{\ell_a} \frac{\partial}{\partial\theta}\Delta_j^\theta & \ell = \ell_a \\ \frac{\partial}{\partial\theta}\Delta_\ell^\theta & \ell_a < \ell < N \\ -\frac{\partial}{\partial\theta}T_N^\theta = -\sum_{j=0}^{N-1} \frac{\partial}{\partial\theta}\Delta_j^\theta & \ell = N \end{cases}, \quad (21)$$

which can all be easily computed during numerical simulation.

The derivations above lead to the following algorithm for the generation of Z_θ over the interval $[0, b]$ and of the random variable $\frac{\partial}{\partial\theta}L_Z(\theta) = \frac{\partial}{\partial\theta} \int_a^b F(\theta, Z_\theta(s)) ds$. The notation in the algorithm provided below is the same as that above with the following exceptions:

- i.) *flag* is a variable that only takes the values zero or one. It starts at zero and becomes one once $t \geq a$. In the algorithm, this moment is determined by finding the first time at which the process makes a jump at a time greater than a (see Step 4 below).
- ii.) The output $\frac{\partial}{\partial\theta}L_Z(\theta)$, as given in (17), is denoted by dL .

It may be helpful for the reader to note that steps 1, 2, 5, 6, 8 and 9 make up the usual implementation of the next reaction method [1, 12]. Only steps 3, 4, 7, and 10 are those required for the derivative terms. All uniform random variables generated in the algorithm below are assumed to be mutually independent.

ALGORITHM. Numerical derivation of Z_θ and $\frac{\partial}{\partial\theta}L_Z(\theta) = \frac{\partial}{\partial\theta} \int_a^b F(\theta, Z_\theta(s)) ds$.

Initialize. Given: a continuous time Markov chain with jump directions ζ_k , intensities $\lambda_k(\theta, z)$, and initial condition z_0 . Set $\ell = 0$, $T_0^\theta = 0$, $Z_\theta(T_0^\theta) = z_0$, $\frac{\partial}{\partial\theta}T_0^\theta = 0$, and $dL = 0$. For each $k \in \{1, \dots, K\}$, set $S_k^\theta(T_0^\theta) = 0$, $\frac{\partial}{\partial\theta}S_k^\theta(T_0^\theta) = 0$. Set *flag* = 0. For each $k \in \{1, \dots, K\}$, set $I_k(T_0^\theta) = \ln(1/u_k)$, where $\{u_k\}$ are independent uniform(0, 1) random variables.

Perform the following steps.

1. For all $k \in \{1, \dots, K\}$, calculate $\lambda_k(Z_\theta(T_\ell^\theta))$. Set

$$\Delta_\ell^\theta = \min_k \frac{I_k(T_\ell^\theta) - S_k^\theta(T_\ell^\theta)}{\lambda_k(\theta, Z_\theta(T_\ell^\theta))} \quad \text{and} \quad j = \operatorname{argmin}_k \frac{I_k(T_\ell^\theta) - S_k^\theta(T_\ell^\theta)}{\lambda_k(\theta, Z_\theta(T_\ell^\theta))}.$$

2. If $T_\ell^\theta + \Delta_\ell^\theta > b$, go to Step 10. Otherwise set $T_{\ell+1}^\theta = T_\ell^\theta + \Delta_\ell^\theta$ and continue to Step 3.
3. Set

$$\frac{\partial}{\partial\theta}\Delta_\ell^\theta = -\frac{\Delta_\ell^\theta}{\lambda_j(\theta, Z_\theta(T_\ell^\theta))} \cdot \frac{\partial}{\partial\theta}\lambda_j(\theta, Z_\theta(T_\ell^\theta)) - \frac{\frac{\partial}{\partial\theta}S_j^\theta(T_\ell^\theta)}{\lambda_j(\theta, Z_\theta(T_\ell^\theta))},$$

then set $\frac{\partial}{\partial\theta}T_{\ell+1}^\theta = \frac{\partial}{\partial\theta}T_\ell^\theta + \frac{\partial}{\partial\theta}\Delta_\ell^\theta$.

4. Set

$$dL \leftarrow dL + \Delta_\ell^\theta \cdot \frac{\partial}{\partial\theta}F(\theta, Z_\theta(T_\ell^\theta)) + F(\theta, Z_\theta(T_\ell^\theta)) \cdot A,$$

where

$$A = \begin{cases} 0 & \text{if } T_{\ell+1}^\theta < a \\ \frac{\partial}{\partial\theta}T_{\ell+1}^\theta & \text{if } T_{\ell+1}^\theta > a \quad \text{and} \quad \text{flag} = 0. \\ \frac{\partial}{\partial\theta}\Delta_\ell^\theta & \text{otherwise} \end{cases}.$$

If $T_{\ell+1}^\theta > a$ and $flag = 0$, set $flag = 1$.

5. Set $Z_\theta(T_{\ell+1}^\theta) = Z_\theta(T_\ell^\theta) + \zeta_j$.

6. For each $k \in \{1, \dots, K\}$, set $S_k^\theta(T_{\ell+1}^\theta) = S_k^\theta(T_\ell^\theta) + \Delta_\ell^\theta \lambda_k(\theta, Z_\theta(T_\ell^\theta))$.

7. For each $k \in \{1, \dots, K\}$, set

$$\frac{\partial}{\partial \theta} S_k^\theta(T_{\ell+1}^\theta) = \frac{\partial}{\partial \theta} S_k^\theta(T_\ell^\theta) + \Delta_\ell^\theta \cdot \frac{\partial}{\partial \theta} \lambda_k(\theta, Z_\theta(T_\ell^\theta)) + \lambda_k(\theta, Z_\theta(T_\ell^\theta)) \cdot \frac{\partial}{\partial \theta} \Delta_\ell^\theta.$$

8. Set $I_j(T_{\ell+1}^\theta) = I_j(T_\ell^\theta) + \ln\left(\frac{1}{u}\right)$, where u is a $\text{uniform}(0, 1)$ random variable.

9. Set $\ell \leftarrow \ell + 1$ and return to Step 1.

10. Set $dL \leftarrow dL + (b - T_\ell^\theta) \frac{\partial}{\partial \theta} F(\theta, Z_\theta(T_\ell^\theta)) - flag \cdot F(\theta, Z_\theta(T_\ell^\theta)) \cdot \frac{\partial}{\partial \theta} T_\ell^\theta$.

2.2.4 Validity of pathwise estimators

Letting Z_θ be a process satisfying a stochastic equation of the form (1), we turn to the question of when $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_Z(\theta)] = \mathbb{E}[\frac{\partial}{\partial \theta_i} L_Z(\theta)]$, with $\frac{\partial}{\partial \theta_i} L_Z(\theta)$ detailed in the previous section. For our proof of Theorem 1, we require a condition on the intensity functions of Z_θ that is more restrictive than Condition 1.

Condition 4. Let $\Theta \subset \mathbb{R}^R$. The functions $\lambda_k : \Theta \times \mathbb{Z}^d \rightarrow \mathbb{R}_{\geq 0}$, $k = 1, \dots, K$, satisfy this condition if each of the following hold.

1. There exist constants Γ_M, Γ' such that for all $k \in \{1, \dots, K\}$ and all $z \in \mathbb{Z}^d$,

$$\sup_{\theta \in \Theta} \sup_{z \in \mathcal{S}} \lambda_k(\theta, z) \leq \Gamma_M \quad \text{and} \quad \sup_{\theta \in \Theta} \sup_{z \in \mathcal{S}} \left| \frac{\partial}{\partial \theta_i} \lambda_k(\theta, z) \right| \leq \Gamma'.$$

2. There exists some constant Γ_m such that for all $k \in \{1, \dots, K\}$ and all $z \in \mathbb{Z}^d$,

$$\sup_{\theta \in \Theta} \lambda_k(\theta, z) \neq 0 \Rightarrow \sup_{\theta \in \Theta} \frac{1}{\lambda_k(\theta, z)} \leq \Gamma_m.$$

The first condition guarantees that the intensities and their θ -derivatives are uniformly bounded above. The second condition stipulates that on those $z \in \mathbb{Z}^d$ at which the rates $\lambda_k(\theta, z)$ are not identically zero on Θ , the rates must be uniformly bounded away from zero.

Theorem 1. Suppose that the process Z_θ satisfies the stochastic equation (1) with λ_k satisfying Conditions 3 and 4 on a neighborhood Θ of θ . Suppose that the function F satisfies Condition 2 on Θ . For some $0 \leq a \leq b < \infty$, let $L_Z(\theta) = \int_a^b F(\theta, Z_\theta(s)) ds$. Then $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_Z(\theta)] = \mathbb{E}\left[\frac{\partial}{\partial \theta_i} L_Z(\theta)\right]$, for all $i \in \{1, \dots, R\}$.

The proof of this theorem is similar to that found in [15] and can be found in Appendix A. We believe that the stringent Condition 4 can be replaced by the more relaxed Condition 1, though this remains open. The stricter Condition 4 plays little role in the methods developed here as it can be incorporated into the definition of the process Z_θ , as will be seen in Section 3. In particular, we note that we will not be requiring that our actual process of interest, X_θ , satisfies Condition 4, only that the approximate process, Z_θ , does.

2.3 Likelihood ratios and coupled paths

The likelihood ratio (LR) method for sensitivity estimation proceeds by selecting a realization of a given process according to a θ -dependent probability measure. Differentiation of the probability measure is then carried out within the expectation. For CTMC models X_θ as in (1) that have θ -differentiable intensities and that satisfy the growth Condition 1 (which, recall, is nearly all biochemical systems), and for a large class of functionals f we have

$$\frac{\partial}{\partial \theta_i} \mathbb{E}[f(\theta, X_\theta)] = \mathbb{E}\left[\frac{\partial}{\partial \theta_i} f(\theta, X_\theta) + f(\theta, X_\theta) H_i(\theta, T)\right] \quad (22)$$

where

$$H_i(\theta, T) = \sum_{\ell=0}^{N(T)-1} \frac{\frac{\partial}{\partial \theta_i} \lambda_{k_\ell}(\theta, \hat{X}_\ell(\theta))}{\lambda_{k_\ell}(\theta, \hat{X}_\ell(\theta))} - \sum_{k=1}^K \int_0^T \frac{\partial}{\partial \theta_i} \lambda_k(\theta, X_\theta(s)) ds, \quad (23)$$

and where

- $N(T)$ is the total number of jumps of X_θ through time T , and a sum of the form $\sum_{\ell=0}^{-1}$ is set to zero,
- k_ℓ is the index of the reaction that changes the system from the ℓ th state to the $(\ell + 1)$ st state,
- $\hat{X}_\ell(\theta)$ is the ℓ^{th} state in the embedded discrete chain of the path.

For a system (1) with intensities of the form $\lambda_k(\theta, x) = \theta_k g_k(x)$, where $g_k : \mathbb{Z}^d \rightarrow \mathbb{R}_{\geq 0}$, such as stochastic mass action kinetics, H_i simplifies to

$$H_i(\theta, T) = \frac{1}{\theta_i} \left(N_i(T) - \int_0^T \lambda_i(\theta, X_\theta(s)) ds \right)$$

where $N_i(T)$ is the number of jumps of reaction i by time T . See [8, 16, 26].

The random variable $H_i(\theta, T)$ is often known as a weighting function or weight, and is simple to compute during path simulation. The likelihood ratio method is widely applicable, straightforward to use, and provides an unbiased estimate of the sensitivity. However, the variance of the estimate is often prohibitively large, leading to an inefficient method. One can reduce this variance significantly by using the weight as a control variate (see e.g. Section V.2 of [8]), since $H_i(\theta, \cdot)$ is often a mean zero martingale [6].

2.3.1 The LR method applied to coupled paths

As was pointed out in and around (8), we want to apply the likelihood ratio method to estimate the sensitivity $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_X(\theta) - L_Z(\theta)]$ where X and Z are coupled processes. Assume that X_θ and Z_θ have the same jump directions $\zeta_k \in \mathbb{Z}^d$, but different intensity functions. Denote their respective intensity functions by λ_k^X and λ_k^Z . It may happen that X_θ and Z_θ have different natural state spaces. In particular, the most common application will have $X_\theta(t) \in \mathbb{Z}_{\geq 0}^d$ while $Z_\theta(t) \in \mathbb{Z}^d$. Therefore, we simply take the domains of λ_k^X and λ_k^Z to be the union of the two; for example, all of \mathbb{Z}^d . If the natural domain of either intensity function is some subset of \mathbb{Z}^d , then that function will need to be extended to this larger domain in some reasonable fashion. For example, since the natural domain of λ_k^X is often $\mathbb{Z}_{\geq 0}^d$, we may extend each λ_k^X to be identically zero outside of the non-negative orthant.

To proceed we must couple the process X_θ and Z_θ ; i.e. we must build them on the same probability space. We will use the split coupling, which first appeared in [23] and has since appeared in numerous publications related to computational methods [2, 3, 4, 5, 18, 33]. We take

$$W_\theta(t) := \begin{bmatrix} X_\theta(t) \\ Z_\theta(t) \end{bmatrix}$$

to be the family of processes satisfying the stochastic equation

$$\begin{aligned} X_\theta(t) &= X_\theta(0) + \sum_{k=1}^K Y_{k,1} \left(\int_0^t \lambda_k^X(\theta, X_\theta(s)) \wedge \lambda_k^Z(\theta, Z_\theta(s)) ds \right) \zeta_k \\ &\quad + Y_{k,2} \left(\int_0^t \lambda_k^X(\theta, X_\theta(s)) - \lambda_k^X(\theta, X_\theta(s)) \wedge \lambda_k^Z(\theta, Z_\theta(s)) ds \right) \zeta_k, \\ Z_\theta(t) &= Z_\theta(0) + \sum_{k=1}^K Y_{k,1} \left(\int_0^t \lambda_k^X(\theta, X_\theta(s)) \wedge \lambda_k^Z(\theta, Z_\theta(s)) ds \right) \zeta_k \\ &\quad + Y_{k,3} \left(\int_0^t \lambda_k^Z(\theta, Z_\theta(s)) - \lambda_k^X(\theta, X_\theta(s)) \wedge \lambda_k^Z(\theta, Z_\theta(s)) ds \right) \zeta_k, \end{aligned} \quad (24)$$

where $\{Y_{k,1}, Y_{k,2}, Y_{k,3}\}$ are independent unit-rate Poisson processes and we recall that $a \wedge b = \min(a, b)$ for any $a, b \in \mathbb{R}$. Note that the $2d$ -dimensional process $W_\theta(t)$ is also a CTMC. For each $k \in \{1, \dots, K\}$ the reaction of the system (1) with reaction vector $\zeta_k \in \mathbb{Z}^d$ has been associated with three reactions of the process W_θ . The reaction vectors for these three reactions, which are elements of \mathbb{Z}^{2d} , are

$$\eta_{k,1} = \begin{bmatrix} \zeta_k \\ \zeta_k \end{bmatrix}, \quad \eta_{k,2} = \begin{bmatrix} \zeta_k \\ 0 \end{bmatrix}, \quad \eta_{k,3} = \begin{bmatrix} 0 \\ \zeta_k \end{bmatrix},$$

where each 0 is interpreted as $\vec{0} \in \mathbb{Z}^d$. Letting $w = \begin{pmatrix} x \\ z \end{pmatrix} \in \mathbb{Z}^{2d}$, where $x, z \in \mathbb{Z}^d$, the intensity functions for the three reactions are

$$\begin{aligned} \Lambda_{k,1}(\theta, w) &= \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z), \\ \Lambda_{k,2}(\theta, w) &= \lambda_k^X(\theta, x) - \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z), \\ \Lambda_{k,3}(\theta, w) &= \lambda_k^Z(\theta, z) - \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z). \end{aligned} \tag{25}$$

We say a reaction associated with W_θ is of *type* $j \in \{1, 2, 3\}$ if the reaction vector is $\eta_{k,j}$. Now note that

$$W_\theta(t) = W_\theta(0) + \sum_{j=1}^3 \sum_{k=1}^K Y_{k,j} \left(\int_0^t \Lambda_{k,j}(\theta, W_\theta(s)) ds \right) \eta_{k,j}$$

has the same general form as (1). Thus, as long as the rates satisfy the usual mild regularity conditions, we may use the likelihood method as in (22)–(23). Given some function $\tilde{f} : \mathbb{R}^R \times D_{\mathbb{Z}^{2d}}[0, \infty) \rightarrow \mathbb{R}$, the analogous equations are

$$\frac{\partial}{\partial \theta_i} \mathbb{E}[\tilde{f}(\theta, W_\theta)] = \mathbb{E} \left[\frac{\partial}{\partial \theta_i} \tilde{f}(\theta, W_\theta) + \tilde{f}(\theta, W_\theta) \tilde{H}_i(\theta, T) \right] \tag{26}$$

where

$$\tilde{H}_i(\theta, T) = \sum_{\ell=0}^{\tilde{N}(T)-1} \frac{\frac{\partial}{\partial \theta_i} \Lambda_{k_\ell, j_\ell}(\theta, \hat{W}_\ell(\theta))}{\Lambda_{k_\ell, j_\ell}(\theta, \hat{W}_\ell(\theta))} - \sum_{j=1}^3 \sum_{k=1}^K \int_0^t \frac{\partial}{\partial \theta_i} \Lambda_{k,j}(\theta, W_\theta(s)) ds,$$

and where

- $\tilde{N}(T)$ is the total number of jumps of $W(\theta)$ through time T ,
- $k_\ell \in \{1, \dots, K\}$ is the index and $j_\ell \in \{1, 2, 3\}$ is the type of the reaction that changes W_θ from the ℓ th state to the $(\ell + 1)$ st state,
- $\hat{W}_\ell(\theta)$ is the ℓ^{th} state in the embedded discrete chain of the path of W_θ , with enumeration starting at $\ell = 0$.

For a system in which $\Lambda_{i,j}(\theta, w) = \theta_k g_{i,j}(w)$, \tilde{H}_i simplifies to

$$\tilde{H}_i(\theta, T) = \sum_{j=1}^3 \left[\frac{1}{\theta_i} \left(\tilde{N}_{i,j}(T) - \int_0^T \Lambda_{i,j}(\theta, W_\theta(s)) ds \right) \right],$$

where $\tilde{N}_{i,j}(T)$ is the number of jumps of reaction i of type j by time T .

We return to our problem at hand of estimating

$$\frac{\partial}{\partial \theta_i} \mathbb{E}[L_X(\theta) - L_Z(\theta)] = \frac{\partial}{\partial \theta_i} \mathbb{E} \left[\int_a^b F(\theta, X_\theta(s)) - F(\theta, Z_\theta(s)) ds \right].$$

Using (26) with $\tilde{f}(\theta, W_\theta) = \int_a^b [F(\theta, X_\theta(s)) - F(\theta, Z_\theta(s))] ds$, we see that, so long as the differentiation is valid, $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_X(\theta) - L_Z(\theta)] = \mathbb{E}[V(\theta)]$ with

$$V(\theta) := \int_a^b \left(\frac{\partial}{\partial \theta_i} F(\theta, X_\theta(s)) - \frac{\partial}{\partial \theta_i} F(\theta, Z_\theta(s)) \right) ds + \tilde{H}_i(\theta, b) \int_a^b [F(\theta, X_\theta(s)) - F(\theta, Z_\theta(s))] ds, \tag{27}$$

where the partial of F is always with respect to the first variable.

2.3.2 Requirements for the process Z_θ .

So long as the rates of both X_θ and Z_θ are differentiable, the new rates (25) for the coupled process are piecewise differentiable. However, because the intensities $\Lambda_{k,j}$ involve minima, there may be values of θ and w where the derivative does not exist. In particular, this may occur if, for some k , the two rates in the minimum $\lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z)$ are equal, since at such points the left- and right-hand derivatives may be different.

The following condition ensures the differentiability of each $\Lambda_{k,j}$.

Condition 5. Suppose for some $k \in \{1, \dots, K\}$ and some $w = \begin{pmatrix} x \\ z \end{pmatrix}$ in the state space of W we have that $\lambda_k^X(\theta, x) = \lambda_k^Z(\theta, z)$. Then we require that $\frac{\partial}{\partial \theta_i} \lambda_k^Z(\theta, z) = \frac{\partial}{\partial \theta_i} \lambda_k^X(\theta, x)$ for each $i \in \{1, \dots, R\}$.

3 The hybrid pathwise method

3.1 Putting it all together

Developing hybrid pathwise methods is now straightforward. We will estimate $\nabla_\theta \mathbb{E}[L_X(\theta)]$ using (8) for an appropriately chosen process Z_θ . In Section 2, we detailed the main conditions that Z_θ must satisfy for this procedure to work. Specifically, we need a Z_θ that is tightly coupled with X_θ , that satisfies the non-interruptive Condition 3, and that satisfies the regularity Conditions 4 and 5. We also require that F , which determines L via (6), satisfies Condition 2. Finally, for the validity of the likelihood ratio method on the error term, we require that X_θ satisfies Condition 1. The hybrid pathwise method then proceeds by

1. estimating $\nabla_\theta \mathbb{E}[L_X(\theta) - L_Z(\theta)]$ via Monte Carlo using the LR method as detailed in Section 2.3.1, and
2. estimating $\nabla_\theta \mathbb{E}[L_Z(\theta)]$ via Monte Carlo using the pathwise method as detailed in Section 2.2.3.

Denoting by Q_{X-Z} and Q_Z the two estimators detailed above, our final estimate for $\nabla_\theta \mathbb{E}[L_X(\theta)]$ is taken to be

$$Q_X := Q_{X-Z} + Q_Z, \quad (28)$$

which is trivially unbiased. We will generate paths independently, in which case

$$\text{Var}(Q_X) = \text{Var}(Q_{X-Z}) + \text{Var}(Q_Z), \quad (29)$$

which can be estimated and used for confidence intervals in the usual way.

Any Z_θ satisfying the above conditions may be used. In order to make specific suggestions, we now restrict ourselves to the setting of biochemistry where, as detailed in the introduction, $\zeta_k = \nu'_k - \nu_k$ and the natural state space of X_θ is $\mathbb{Z}_{\geq 0}^d$. We will consider two cases: when λ_k^X satisfies stochastic mass action kinetics and when λ_k^X satisfies Michaelis–Menten kinetics.

Stochastic mass action kinetics. Suppose that $\lambda_k^X(\theta, x)$ satisfies stochastic mass action kinetics (2), in which case $\lambda_k^X(\theta, x) = \theta_k g_k(x)$. We define $\lambda_k^X(\theta, x) = 0$ if $x \notin \mathbb{Z}_{\geq 0}^d$.

We now define Z_θ to be the process satisfying (1) with the following intensity functions. For each $k \in \{1, \dots, K\}$ let $\delta_k > 0$. Let $M > 0$ be a large number. Define

$$\lambda_k^Z(\theta, z) = \begin{cases} \theta_k \delta_k & \text{if } z_i < \nu_{ki} \text{ for any } i \text{ such that } \nu_{ki} > 0 \\ \theta_k M & \text{if } \lambda_k^X(\theta, z) \geq \theta_k M \\ \lambda_k^X(\theta, z) & \text{otherwise} \end{cases}. \quad (30)$$

Note that in much of $\mathbb{Z}_{\geq 0}^d$ the rates of Z_θ are identical to those of X_θ . Note also that Z_θ satisfies the non-interruptive Condition 3, the restrictive regularity Condition 4, and the Condition 5 guaranteeing the applicability of the LR method on the coupled processes. The redefinition of the intensity functions for large values of $\lambda_k^X(\theta, z)$ (by $\theta_k M$) is a consequence of our theoretical results. If Theorem 1 can be proven with Condition 4 replaced by Condition 1, as we believe is possible, then the M term could be ignored and we would have

$$\lambda_k^Z(\theta, z) = \begin{cases} \theta_k \delta_k & \text{if } z_i < \nu_{ki} \text{ for any } i \text{ such that } \nu_{ki} > 0 \\ \lambda_k^X(\theta, z) & \text{otherwise} \end{cases}.$$

Michaelis–Menten Kinetics. The standard Michaelis–Menten rate is of the form $\lambda_k^X(\theta, x) = \frac{\theta_1 x_k}{\theta_2 + x_k}$ [29]. Note that near a fixed θ this rate is uniformly bounded in $x \geq 0$. For some $\delta_k > 0$ let

$$\lambda_k^Z(\theta, z) = \begin{cases} \frac{\theta_1 \delta_k}{\theta_2 + \delta_k} & \text{if } z_i < \nu_{ki} \text{ for any } i \text{ such that } \nu_{ki} > 0 \\ \lambda_k^X(\theta, z) & \text{otherwise.} \end{cases} \quad (31)$$

Note that (i) Z_θ so defined will again have rates that are in agreement with X_θ for much of $\mathbb{Z}_{\geq 0}^d$, and (ii) Z_θ satisfies all the conditions outlined above, including the non-interruptive Condition 3.

It is important to note that the processes Z_θ defined in the manner of (30) or (31) can reach states with negative coordinates, even if the initial condition $Z_\theta(0)$ is in $\mathbb{Z}_{\geq 0}^d$. This is a consequence of how we overcame the problem that, in general, biochemical processes do not satisfy the non interruptive condition 3.

3.2 Implementation issues

In this short section, we make a few points about implementing the hybrid pathwise method.

1. In the previous section, we were conservative in redefining *all* intensity functions so that they can never become zero. However, if a reaction cannot be interrupted by another, then there is no need to redefine the kinetics at zero. Allowing such intensities to become zero will then improve the performance of the method. For example, see the model in Section 4.2.

In particular, if the process X_θ already satisfies the non-interruptive Condition 3 and the restrictive Condition 4, then the approximate process Z_θ is unnecessary: one can use pathwise estimates alone to estimate $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_X(\theta)]$. See Section 4.1 for such an example.

2. The best choice for the δ_k of (30) and (31) will be model-dependent. If δ_k is too large, the process Z_θ may cease to be a good approximation of X_θ , which will cause the variance of the likelihood ratio estimate of $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_X(\theta) - L_Z(\theta)]$ to be large. On the other hand, taking δ_k too small makes it very rare that the process Z_θ makes a jump that the process X_θ cannot make. In this latter case, the problem of estimating $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_X(\theta) - L_Z(\theta)]$ becomes a problem of estimating a rare event.

In our numerical experiments, we found that taking δ_k to be near one was a reasonable choice for all the models we considered. Additionally, we have found that M can be taken arbitrarily large with no loss of accuracy.

3. If the sensitivity we wish to estimate is of the form $\frac{\partial}{\partial \theta_i} \mathbb{E}[f(X_\theta(T))]$, i.e. is not an integral of a function, we may instead write

$$\frac{\partial}{\partial \theta_i} \mathbb{E}[f(X_\theta(T))] = \frac{\partial}{\partial \theta_i} \mathbb{E}[f(X_\theta(T)) - f(Z_\theta(T))] + \frac{\partial}{\partial \theta_i} \mathbb{E}[f(Z_\theta(T))], \quad (32)$$

and note that the LR method is applicable on the first term on the right-hand side of the above equation. That is, there is no reason to replace f in that term with an integrated function. The final term must be estimated using either the GS smoothing method or the RPD smoothing method. We shall refer to these hybrid procedures for estimating $\frac{\partial}{\partial \theta_i} \mathbb{E}[f(X_\theta(T))]$ as the **GS hybrid** and the **RPD hybrid** methods, respectively.

4. One must decide how many simulated paths will be used for each of the estimators Q_{X-Z} and Q_Z of (28). Suppose one wishes to minimize the expected time required to compute an estimate such that the 95% halfwidth is within some target value, ϵ . That is, we would like

$$\text{Var}(Q_{X-Z}) + \text{Var}(Q_Z) = \text{Var}(Q_X) \leq \delta := \left(\frac{\epsilon}{1.96} \right)^2, \quad (33)$$

where δ denotes the target variance. Let v_ℓ denote the variance $\text{Var}(V(\theta))$, where $V(\theta)$ is as in (27), so that $\text{Var}(Q_{X-Z}) = \frac{v_\ell}{n_\ell}$, where n_ℓ is the number of coupled paths simulated. Also let c_ℓ denote the average time required to compute one pair of coupled paths for the likelihood estimate. Similarly define

v_p , c_p , and n_p for the pathwise estimates. Then, we wish to minimize the expected total computational time

$$n_\ell c_\ell + n_p c_p = \frac{v_\ell c_\ell}{\text{Var}(Q_{X-Z})} + \frac{v_p c_p}{\text{Var}(Q_Z)}$$

subject to the constraint (33). The solution to this optimization problem satisfies

$$\text{Var}(Q_{X-Z}) = \frac{\delta \sqrt{v_\ell c_\ell}}{\sqrt{v_p c_p} + \sqrt{v_\ell c_\ell}} \quad \text{and} \quad \text{Var}(Q_Z) = \frac{\delta \sqrt{v_p c_p}}{\sqrt{v_p c_p} + \sqrt{v_\ell c_\ell}}. \quad (34)$$

In practice, one may use the following optimization procedure. First, in a preliminary simulation compute n samples each of (X_θ, Z_θ) and Z_θ . Second, from these preliminary samples estimate each of v_p, c_p, v_ℓ , and c_ℓ and utilize these values to estimate the target variances (34).

5. Finally, we point out that if one first simulates many paths of Z_θ for use in the pathwise estimate Q_Z and notes that each path is a valid realization of the original process X_θ (which is simple to check as simulation occurs), then with high probability one knows without further computation that Q_{X-Z} is zero or near zero. Of course, theoretical work is needed to quantify what is meant by “high probability” in the previous sentence. However, this observation provides a means to check for practical applicability of pathwise methods, which have been shown to be extremely efficient on many models [30].

4 Numerical Examples

With the examples in this section, we demonstrate the validity and efficiency of our new class of methods. An important example is given in Section 4.2, where we demonstrate that pathwise-only methods of the type developed in [30] can fail, in the sense that there can be large biases, if interruptions can occur. That is, the example in Section 4.2 shows that the error term utilized in this paper, and differentiated using the LR method, is necessary.

On a variety of examples we compare the efficiency of the developed methods with the following:

1. The likelihood ratio method including the weight (23) as a control variate (LR+CV).
2. The regularized pathwise derivative method (RPD).
3. The coupled finite difference method (CFD) using centered differences.
4. The Poisson path approximation method (PPA).

We will demonstrate that the new methods introduced here compare quite favorably with this group of already established methods, with the GS hybrid method often the most efficient unbiased method. Future work will involve a wider numerical study to help determine a better framework for choosing the most efficient method for a given model.

Throughout, we use the term “variance” to refer to estimator variance, which is the sample variance divided by the number of paths simulated. For each hybrid method estimate, we use the optimization procedure described in item 4 of Section 3.2, and compute the variance as in (29). All half-widths given are 95% confidence intervals computed as 1.96 multiplied by the square root of the variance. The numerical calculations were carried out in MATLAB using an Intel i5-4570 3.2 GHz quad-core processor.

4.1 Birth-death

Consider the birth-death model



with mass action kinetics. We let $X_\theta(t)$ denote the abundance of A at time t and take $X_\theta(0) = 0$. For this model, we can solve to find that

$$\mathbb{E}[X_\theta(t)] = \frac{\theta_1}{\theta_2}(1 - e^{-\theta_2 t}),$$

Comparison of θ_2 sensitivity estimates, Birth-Death model

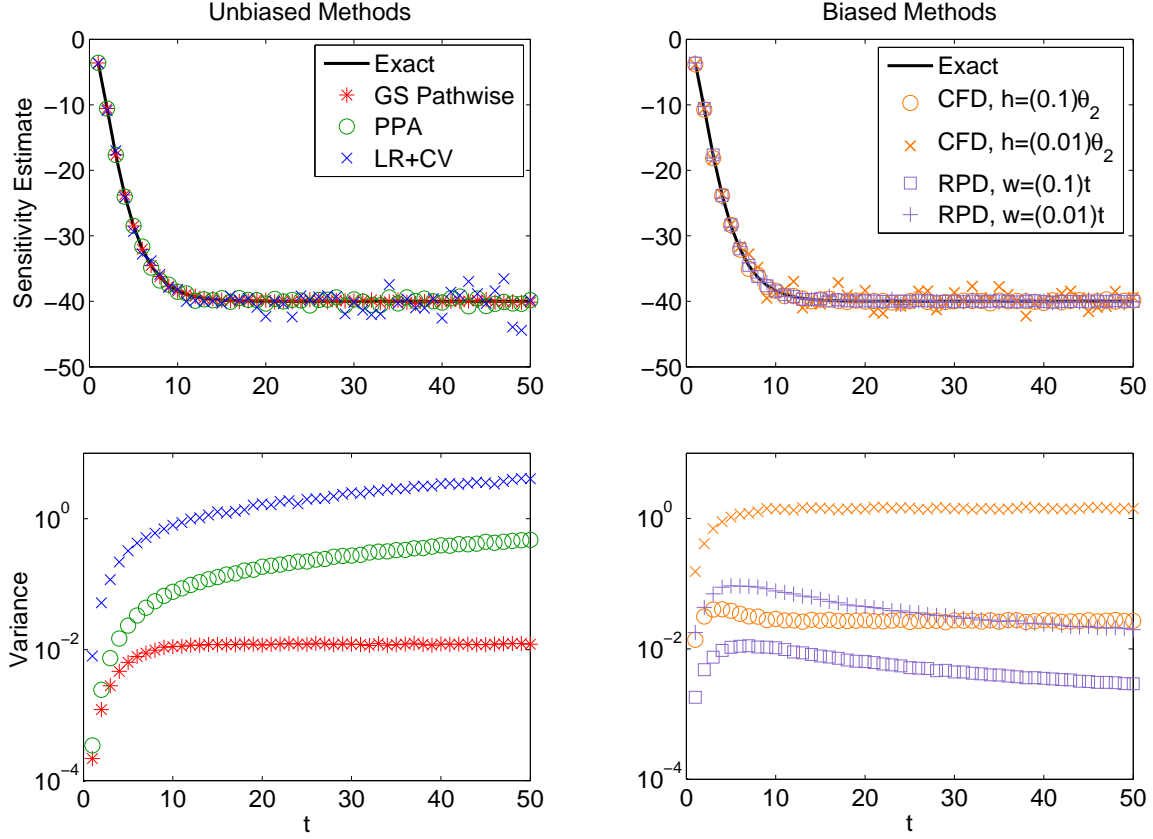


Figure 1: A comparison of the sensitivity estimates and method variance for the birth-death model of Section 4.1 as the time t is varied. 10^4 paths were used for each method, and $X_\theta(0) = 0$, and $\theta_0 = (10, 0.5)$. The parameter h for the CFD method was chosen as a fraction of the parameter θ_2 . Similarly, the parameter w for the RPD method was chosen as a fraction of the time t , which varies in this experiment.

and

$$\frac{\partial}{\partial \theta_1} \mathbb{E}[X_\theta(t)] = \frac{1}{\theta_2} (1 - e^{-\theta_2 t}) \quad \text{and} \quad \frac{\partial}{\partial \theta_2} \mathbb{E}[X_\theta(t)] = \frac{\theta_1}{\theta_2} (te^{-\theta_2 t}) - \frac{\theta_1}{\theta_2^2} (1 - e^{-\theta_2 t}).$$

We estimate the sensitivity with respect to θ_2 of the quantity $\mathbb{E}[X_\theta(t)]$ at $\theta_0 = (\theta_1, \theta_2) = (10, 0.5)$.

Since the model naturally satisfies Condition 3 we may use the GS Pathwise and RPD methods without the error terms; see item 1 of Section 3.2. Though the intensity of the model is unbounded, the intensities are “bounded in practice:” throughout these simulations no intensity was ever greater than $M = 10^3$. That is, if we had used the full hybrid method with an approximate process Z_θ with an intensity bounded above by 10^3 , then the error term would have given us an estimate of zero. We may therefore confidently use both pathwise-only methods.

Figure 1 shows that each method does a good job of estimating the given sensitivities and that the GS pathwise method has the lowest variance of any unbiased method. In fact, for this experiment the GS pathwise method also has a smaller variance than most of the biased methods. The RPD method, with the larger choice of w , has a slightly lower variance than the GS pathwise method.

A more straightforward comparison of method efficiency can be provided by finding the CPU time required for each method to estimate the sensitivity to a given tolerance. In Figure 2, we report these CPU times when we run each method until it produced a half-width equal to 1% of the absolute value of the sensitivity. As can be seen in the figure, the GS pathwise method is significantly more efficient than the other unbiased methods. Indeed, at time $t = 5$, the GS pathwise method is over 3 times faster than PPA, and more than 20 times faster than the LR+CV method. At time $t = 50$, the GS pathwise method is over 25 times faster

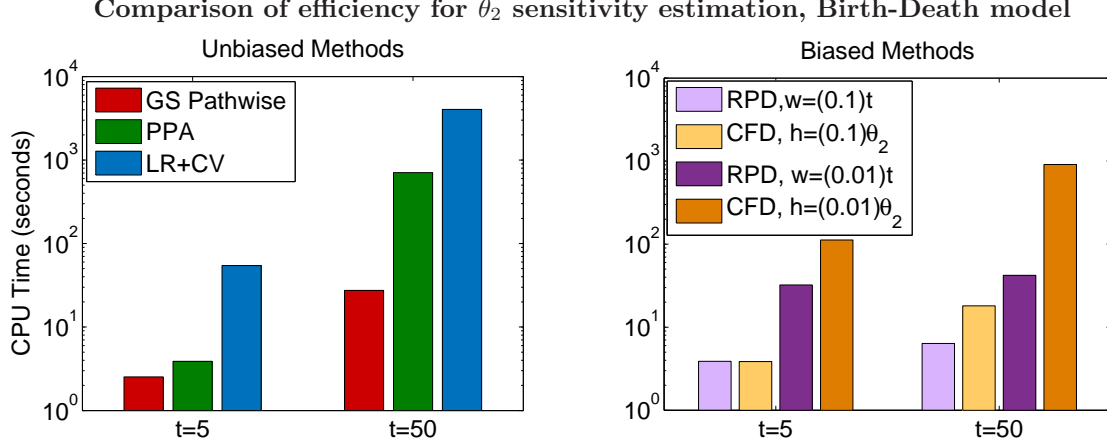


Figure 2: A comparison of the efficiency of the different methods in estimating the sensitivity with respect to θ_2 of the birth-death model of Section 4.1 with $X_\theta(0) = 0$, and $\theta_0 = (10, 0.5)$. Two different times, 5 and 50, were used. The CPU times reported are the times required by the different methods to produce a target confidence interval of half-width equal to 1% of the absolute value of the sensitivity. Note that a log scale is used.

than PPA, and nearly 150 times faster than the LR+CV method. At time $t = 5$, the GS pathwise method is also more efficient than any of the biased methods used.

The efficiency of the biased methods RPD and CFD is highly influenced by the choice of the parameter w or h . At time $t = 50$, the RPD method with $w = (0.1)t = 5$ is over 4 times faster than the GS pathwise method, though at the cost of a small bias.

Finally, we note here that on this model and the other models simulated, the LR+CV method, which uses the weighting function as a control variate, generally has variances at least an order of magnitude smaller than the usual LR method in which a control variate is not used. The additional computational cost of adding this control variate is negligible.

4.2 A simple switch

In contrast to the linear growth model, the following simple switch is one in which the two pathwise-only methods can have a large bias if no correction term is added:

$$A \xrightarrow{\theta_1} \emptyset, \quad A \xrightarrow{\theta_2} B, \quad B \xrightarrow{\theta_3} C,$$

with $X_\theta(0) = (a, 0, 0)$ giving the initial abundances of A, B , and C respectively. We estimate the derivative with respect to θ_1 of the mean number of C molecules, $\frac{\partial}{\partial \theta_1} \mathbb{E}[X_{\theta,C}(t)]$, at $\theta = (\frac{1}{4}, 1, 1)$ and at various times t . Since this model is linear, we can solve for the sensitivity exactly at $\theta = (\theta_1, 1, 1)$:

$$\frac{\partial}{\partial \theta_1} \mathbb{E}[X_{\theta,C}(t)] = \frac{ae^{-t}}{\theta_1^2} - \frac{a}{(1+\theta_1)^2} - ae^{-(1+\theta_1)t} \left(\frac{\theta_1^2 t + \theta(t+2) + 1}{\theta_1^2(1+\theta_1)^2} \right).$$

4.2.1 Pathwise-only methods are biased

We consider the bias of the GS pathwise and RPD methods in computing the sensitivity $\frac{\partial}{\partial \theta_1} \mathbb{E}[X_{\theta,C}(t)]$. For the GS pathwise method we use $\mathbb{E}[X_{\theta,C}(t)] = \mathbb{E}[\int_0^t X_{\theta,B}(s) ds]$, which follows from (12). For the RPD method, we use

$$\mathbb{E} \left[\frac{1}{2w} \int_{T-w}^{T+w} X_{\theta,C}(s) ds \right]$$

as an approximation to $\mathbb{E}[X_{\theta,C}(T)]$. As shown in Figure 3, the RPD and GS pathwise methods provide biased estimates, with the bias ranging from small to (very) large, depending on the initial condition and

time, t . In fact, at $t = 10$, these two methods provide estimates of approximately zero for a sensitivity of magnitude approximately 6. At a small time of $t = 0.5$, the RPD and GS pathwise methods show only a small bias, though it is still noticeable for small initial abundances of A . In each plot of Figure 3, the same value of w was used for both the RPD and RPD Hybrid methods (the hybrid methods for this example are discussed below).

These results confirm that neither the RPD method nor the GS pathwise method is unbiased for models with interruptions. Further, the biases can be substantial.

4.2.2 Comparison of valid methods

To use the hybrid methods introduced in this paper, we construct Z_θ as in Section 3 with

$$\lambda_1^Z(\theta, z) = \begin{cases} \frac{1}{4} & z_A < 1 \\ \frac{1}{4}z_A & \text{otherwise} \end{cases}, \quad \lambda_2^Z(\theta, z) = \begin{cases} 1 & z_A < 1 \\ z_A & \text{otherwise} \end{cases}, \quad \lambda_3^Z(\theta, z) = \begin{cases} 0 & z_B < 1 \\ z_B & \text{otherwise} \end{cases}. \quad (35)$$

The process Z_θ may now reach states in which the first coordinate is negative. We may allow the rate $\lambda_3^Z(\theta, x)$ to be zero because the reaction $B \rightarrow C$ can never be interrupted by another reaction; see item 1 of Section 3.2. Hence, the Z_θ constructed with rates (35) is still non-interruptive.

In Figure 4, we give a comparison of method efficiency with $a = 10$ for times $t = 0.5, t = 2$ and $t = 10$. Again, we give the time required for each method to achieve a confidence interval of half-width equal to 1% of the magnitude of the sensitivity. At $t = 0.5$, the GS Hybrid method is significantly more efficient than any other method; in particular, it is almost 10 times faster than PPA and over 165 times faster than LR+CV, the other unbiased methods considered. As time increases to $t = 2$, however, PPA becomes the most efficient method. At $t = 10$ the advantage of PPA over the hybrid methods is even more significant: PPA is over 30 times faster than the GS Hybrid method. Interestingly, at $t = 10$, the LR+CV method is very nearly as efficient as PPA. This is a particularly striking example of why future work should include a study of the regimes in which a given method is likely to be the most efficient choice.

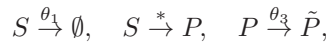
Note that in this example the biased methods with the given parameter choices are less efficient than the most efficient unbiased method at each time we considered.

4.2.3 Michaelis–Menten kinetics

We demonstrate the hybrid methods on a non-mass action model. In particular, the standard Michaelis–Menten approximation of the substrate–enzyme model



would lead to the model



where the intensity $(*)$ is given by $\lambda_2^X(\theta, X_\theta) = \frac{\theta_2 X_{\theta,S}}{\theta_4 + X_{\theta,S}}$, and where $X_{\theta,S}$ denotes the number of substrate molecules. The other two rates follow mass action kinetics. See for example [29], from which we obtained the relevant parameter values, $\theta = (1/20, 1, 1, 11)$. Note that this network is analogous to the switch model above. For the needed approximate model we use

$$\lambda_1^Z(\theta, z) = \begin{cases} \frac{1}{20} & z_S < 1 \\ \frac{1}{20}z_S & \text{otherwise} \end{cases}, \quad \lambda_2^Z(\theta, z) = \begin{cases} \frac{\theta_2}{\theta_4+1} & z_S < 1 \\ \frac{\theta_2 z_S}{\theta_4+z_S} & \text{otherwise} \end{cases}, \quad \text{and} \quad \lambda_3^Z(\theta, z) = \begin{cases} 0 & z_P < 1 \\ \theta_3 z_P & \text{otherwise} \end{cases}.$$

Again note that the third reaction cannot be interrupted. We estimate $\frac{\partial}{\partial \theta_1} \mathbb{E}[X_{\theta, \tilde{P}}(t)]$ at times $t = 2$ and $t = 20$; the actual sensitivity values are approximately 0.23 and 29 respectively. The results are similar to the results of the mass action switch model of Section 4.2.2. See Figure 5. In particular, for the small time $t = 2$, the hybrid methods are more efficient than PPA and the other methods. In particular, the GS Hybrid method is over 7 times faster than PPA. At the time of $t = 20$, when the intensity of each reaction channel in the system is often zero, the PPA and LR+CV methods are most efficient, with PPA returning the desired estimate over 12 times faster than the GS Hybrid method.

Error of pathwise-only methods, switch model

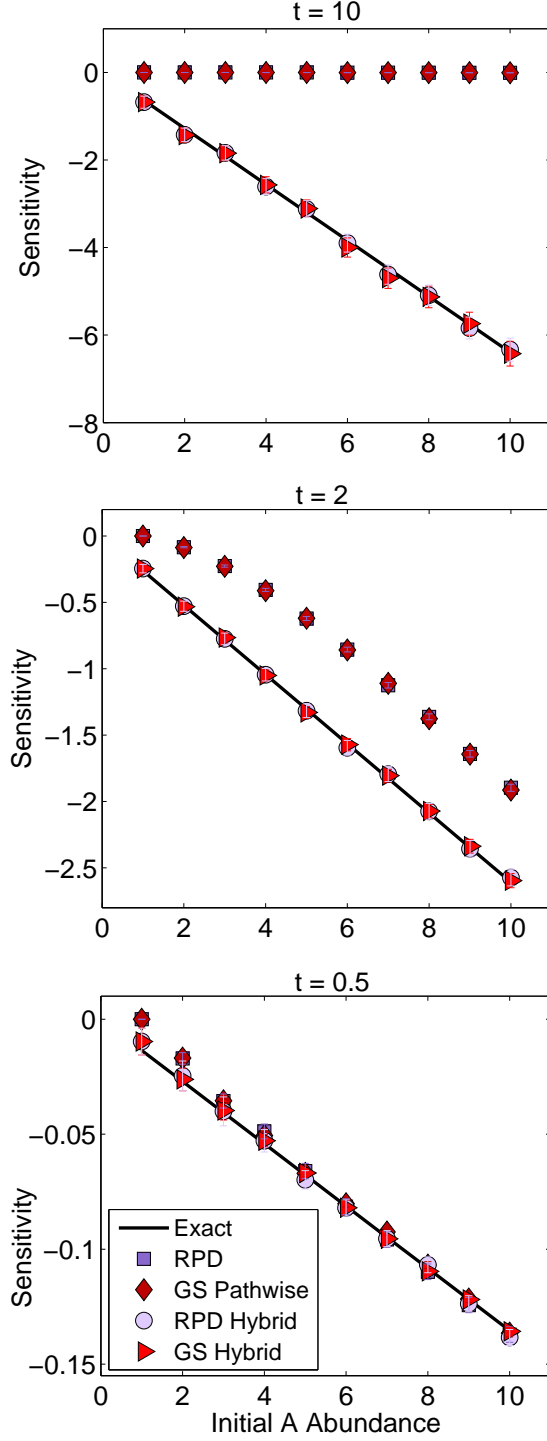


Figure 3: A demonstration of the significant bias of the pathwise-only methods (RPD and GS) for the estimation of the sensitivity of $\mathbb{E}[X_{\theta,C}(t)]$ with respect to θ_1 in the switch model of Section 4.2. Various initial A abundances and three different times t are used. The GS Hybrid and RPD Hybrid method estimates are also shown; both estimate the exact sensitivity well. Each estimate used 10^5 paths, and a value of $w = (0.1)t$ was used for both the RPD and RPD Hybrid methods. For $t = 10$, both hybrid methods used 30% pathwise estimates (and 70% coupled likelihood estimates); at $t = 2$, both used 75% pathwise estimates; and at $t = 0.5$, both used 90% pathwise estimates.

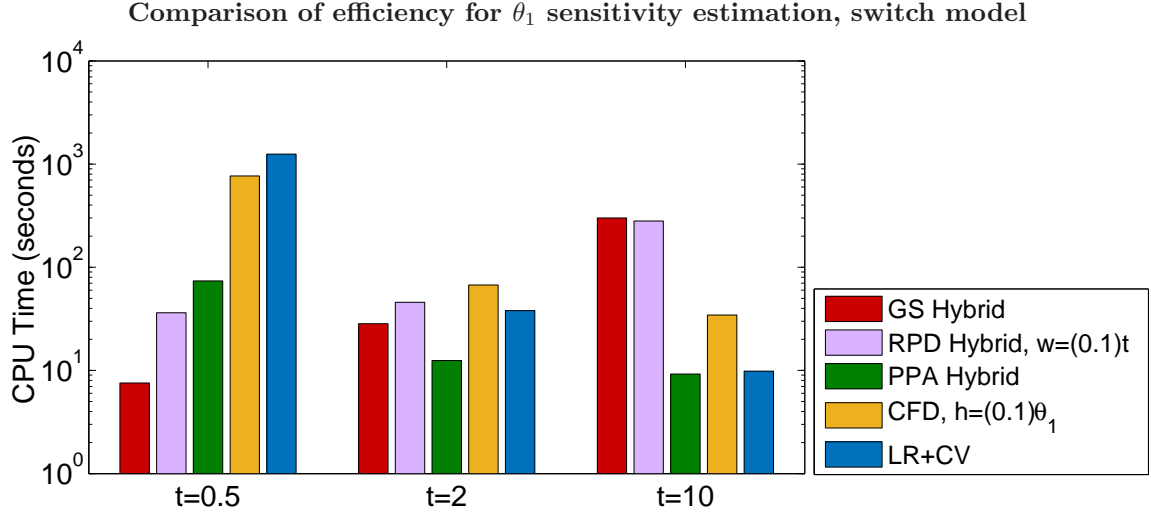


Figure 4: An efficiency comparison for the estimation of the sensitivity of $\mathbb{E}[X_{\theta,C}(t)]$ with respect to θ_1 in the switch model of Section 4.2, with $a = 10$. CPU time gives the computation time in seconds required to achieve a confidence half-width of 1% of the sensitivity. Via the optimization procedure described in Section 3.2, the GS Hybrid method used approximately 36% pathwise estimates, versus 64% coupled likelihood ratio estimates, when $t = 10$; when $t = 2$, the method used 76% pathwise estimates, and when $t = 0.5$ it used 100% pathwise estimates. That is, the best allocation strategy is significantly different at these various times. The RPD Hybrid method similarly uses more pathwise estimates at smaller times, though the exact allocation is different for the two choices of the parameter w . For both hybrid methods, the optimization step is included in the computation time. The time required for the optimization step, which for this experiment included sampling 500 pathwise estimates and 500 coupled likelihood estimates, was approximately 0.10 seconds for $t = 0.5$, 0.15 seconds for $t = 2$, and 0.25 seconds when $t = 10$.

Comparison of efficiency for θ_1 sensitivity estimation, Michaelis-Menten switch model

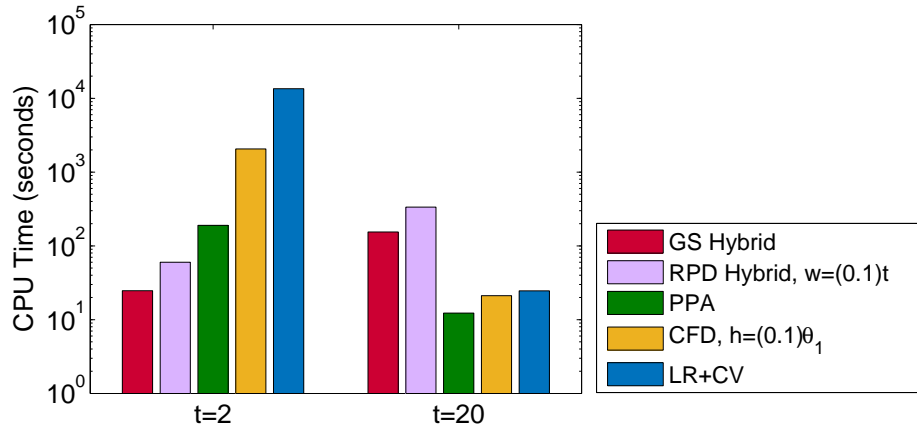


Figure 5: An efficiency comparison for the estimation of $\frac{\partial}{\partial \theta_1} \mathbb{E}[X_{\theta,\bar{P}}(t)]$ in the Michaelis–Menten switch model of Section 4.2.3 with an initial S quantity of 10. CPU time gives computation time in seconds required to achieve a confidence half-width of 1% of the sensitivity value.

4.3 Dimerization

We consider a model of mRNA transcription and translation in which, additionally, the protein dimerizes. Table 1 gives the reactions of the model. Since the model does not satisfy the non-interruptive Condition 3, this table also provides the rates that were used for the approximate process Z_θ in the hybrid methods.

	Reaction	λ_k^X	λ_k^Z
1.) transcription	$\emptyset \rightarrow M$	θ_1	θ_1
2.) translation	$M \rightarrow M + P$	$\theta_2 X_M$	$\begin{cases} \theta_2 & Z_M < 1 \\ \theta_2 \tilde{M} & \theta_2 Z_M \geq \theta_2 \tilde{M} \\ \theta_2 Z_M & \text{otherwise} \end{cases}$
3.) dimerization	$P + P \rightarrow D$	$\theta_3 X_P (X_P - 1)$	$\begin{cases} \theta_3 & Z_P < 2 \\ \theta_3 \tilde{M} & Z_P \geq 2 \text{ and } \\ & \theta_3 Z_P (Z_P - 1) \geq \theta_3 \tilde{M} \\ \theta_3 Z_P (Z_P - 1) & \text{otherwise} \end{cases}$
4.) degradation	$M \rightarrow \emptyset$	$\theta_4 X_M$	$\begin{cases} \theta_4 \tilde{M} & \theta_4 Z_M \geq \theta_4 \tilde{M} \\ \theta_4 Z_M & \text{otherwise} \end{cases}$
5.) degradation	$P \rightarrow \emptyset$	$\theta_5 X_P$	$\begin{cases} \theta_5 & Z_P < 1 \\ \theta_5 \tilde{M} & \theta_5 Z_P \geq \theta_5 \tilde{M} \\ \theta_5 Z_P & \text{otherwise} \end{cases}$
6.) degradation	$D \rightarrow \emptyset$	$\theta_6 X_P$	$\begin{cases} \theta_6 \tilde{M} & \theta_6 Z_D \geq \theta_6 \tilde{M} \\ \theta_6 Z_D & \text{otherwise} \end{cases}$

Table 1: Reactions and hybrid rates for the dimerization model of Section 4.3. We take all initial quantities equal to zero and $\tilde{M} = 10^6$ (we have added a tilde to avoid confusion with the symbol for mRNA). For the process Z_θ to be non-interruptive, we need only prevent three of the intensities from being zero: λ_2, λ_3 , and λ_5 . Indeed, λ_1 is constant, and reactions 4 and 6 cannot be interrupted by another reaction.

4.3.1 Dimer abundance sensitivity

We first estimate the sensitivity $\frac{\partial}{\partial \theta_3} \mathbb{E}[X_{\theta,D}(t)]$ at time $t = 1$, with $\theta = (200, 100, 0.1, 25, 1, 1)$, and with zero initial quantities. In the first bar graph in Figure 6, we show the time required by each method to compute an estimate to within 5% of the sensitivity value. The GS Hybrid method is again the most efficient of the unbiased methods, returning the estimate over 8 times faster than PPA and over 600 times faster than the LR+CV method. In this experiment, for the GS Hybrid method to achieve the target variances determined by the optimization procedure, approximately 53% of the estimates samples were pathwise estimates, with the other 47% being coupled likelihood estimates. See Section 3.2.

The CFD method with $h = (0.1)\theta_3$ is seen to be significantly more efficient than the other methods, including the unbiased methods. Of course, the bias of any such finite difference method is generally unknown, which is an issue if high accuracy is a priority. For example, with $h = (0.1)\theta_3$ the CFD method returns an estimate of 145 ± 1 , while the actual sensitivity is ≈ 141 ; that is, the bias is approximately 3% of the sensitivity value. Furthermore, as expected, the variance is inversely proportional to the size of h , and when h is changed to $(0.01)\theta_3$, the CFD method becomes less efficient than all other methods except LR+CV. This illustrates the issue for biased methods that, a priori, one generally does not know which values of h will provide an efficient estimate with acceptable bias. The RPD hybrid method suffers a similar difficulty in the choice of w : one generally cannot know the bias of a particular w without numerical experimentation.

Comparison of efficiency for θ_3 sensitivity estimation, dimerization model

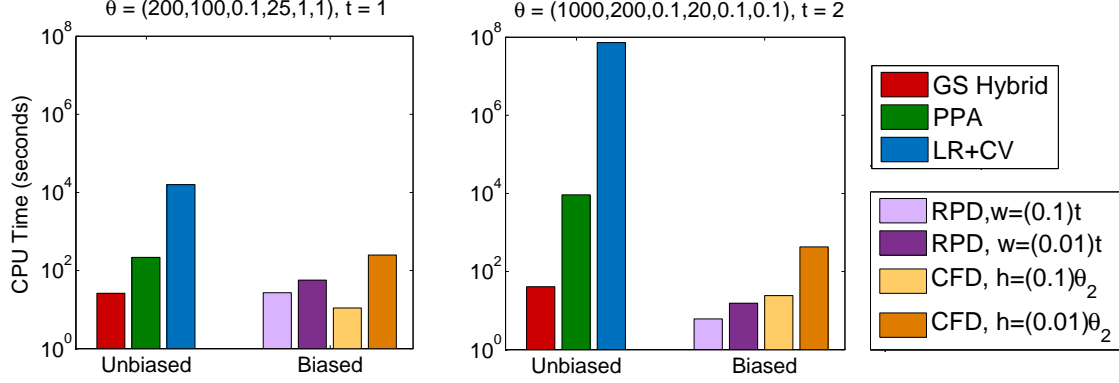


Figure 6: A comparison of efficiency of the sensitivity methods on the dimerization model of Section 4.3 to compute $\frac{\partial}{\partial \theta_3} \mathbb{E}[X_{\theta,D}(t)]$. We provide two estimates. The first estimate is at $\theta = (200, 100, 0.1, 25, 1, 1)$, $t = 1$, and zero initial conditions; the second is at $\theta = (1000, 200, 0.1, 20, 0.1, 0.1)$, $t = 2$, and an initial condition of $X_{\theta,M}(0) = 50$ and other initial abundances equal to 0. CPU gives computation time in seconds required to reach a confidence half-width of 5% of the sensitivity value. In the second graph, the CPU time given for the LR+CV method is an estimate based on the variance of partial data.

For example, with $w = (0.1)t$, the RPD Hybrid method also has a bias of approximately 3%, as it returns an estimate of 145 ± 1 .

We next include results for computing $\frac{\partial}{\partial \theta_3} \mathbb{E}[X_{\theta,D}(t)]$ at a different set of parameters, namely $\theta = (1000, 200, 0.1, 20, 0.1, 0.1)$, at time $t = 2$ and with an initial condition of $X_{\theta,M}(0) = 50$ and other initial abundances equal to 0. As shown in the second graph in Figure 6, in order to achieve a half-width of approximately 5% of the value of the sensitivity, the GS Hybrid method is by far the most efficient unbiased method. In particular, the PPA method requires over 225 times more computation time than the GS hybrid method. We estimate that the LR+CV method requires approximately 1.8×10^6 times more computation time than the GS hybrid method, though we were not able to complete the numerical computations for the LR+CV method due to the fact that the time required to do so was so large. We note that, for this example, the approximate paths Z_θ simulated for the pathwise estimates of the GS Hybrid method were all valid realizations of the original process X_θ . That is, with very high probability, the coupled likelihood estimator is zero or near zero. Thus, contrary to the previous set of parameters, in this experiment, all estimates were pathwise estimates. See Section 3.2.

Note that for this particular experiment, the RPD Hybrid method is more efficient than the GS Hybrid method, by a factor of almost 7 when $w = (0.1)t = 0.2$, and by a factor of about 2.5 when $w = (0.01)t = 0.02$. Furthermore, the bias of the RPD method is less significant than for the previous choice of parameters. In particular, the bias of the RPD Hybrid method when $w = (0.1)t$ is only approximately 1% of the actual value, returning an estimate of 557 ± 1 while the actual value is ≈ 552 ; when $w = (0.01)t$, the bias is only about 0.8%. As described above for the GS Hybrid method, the RPD Hybrid method used only pathwise estimates in this experiment. Also note that the RPD Hybrid method, with either choice of w , is more efficient than the CFD method at either choice of h we considered.

4.3.2 Integrated dimerization rate sensitivity

We consider the functional

$$\int_0^t \lambda_3(\theta, X_\theta(s)) ds = \int_0^t \theta_3 X_{\theta,P}(s) (X_{\theta,P}(s) - 1) ds,$$

	Pathwise hybrid			LR+CV			CFD		
∇_{θ}	0.5713	\pm	0.0067	0.5685	\pm	0.0501	0.5669	\pm	0.0146
	11.48	\pm	0.13	11.14	\pm	0.67	11.26	\pm	0.27
	3401	\pm	34	3162	\pm	308	3403	\pm	126
	-4.559	\pm	0.051	-5.046	\pm	0.419	-4.544	\pm	0.114
	-55.95	\pm	0.59	-57.33	\pm	4.48	-53.32	\pm	1.57
	0.0	\pm	0.0	-0.1	\pm	2.4	0.0	\pm	0.0
CPU Time	68			68			68		

Table 2: A comparison of sensitivity methods on the dimerization model of Section 4.3. Estimates of $\nabla_{\theta}\mathbb{E}[\int_0^t \lambda_5(\theta, X_{\theta}(s)) ds]$ are given for $t = 5$ and at $\theta_0 = (200, 10, 0.01, 25, 1, 1)$. CPU gives computation time in seconds. Recall that the hybrid and LR+CV methods are unbiased, while CFD is not. Note that the total computation time used by each of the three methods is approximately equal (we have rounded the values to the nearest second for clarity). As the CFD method must compute each estimate one by one, the total computation time was allocated approximately equally for each of the six estimates.

which is the integral of the rate of the dimerization reaction, at $t = 5$ and at $\theta_0 = (200, 10, 0.01, 25, 1, 1)$. This quantity is a functional of the path and we therefore use the pathwise hybrid method, outlined in and around (8), on this quantity directly. That is, we do not need to use the martingale representation (11) as we have in previous examples. The RPD and PPA methods are not applicable for the computation of this sensitivity. Also note that, unlike in previous examples, the functional depends explicitly on θ , which requires the methods to take into account the partial derivative of the functional in both pathwise and likelihood ratio estimators.

Instead of estimating a single derivative, we estimate the full gradient. Further, for this example we estimate the efficiency of the methods by simulating each valid method for a fixed amount of time and comparing the resulting confidence intervals for each of the entries of the gradient. Table 2 provides this comparison for the pathwise hybrid, the LR+CV, and the CFD methods. As shown in the table, the pathwise hybrid method is significantly more precise than the LR+CV method, which is the only other unbiased method that is applicable for this problem. The pathwise hybrid method is also significantly more precise than the CFD method, which for this experiment used the relatively large perturbations of $h = (0.1)\theta_i$ for the i th entry of the gradient (which leads to a smaller variance). The relatively poor behavior of the CFD method is partially due to the fact that, unlike the pathwise hybrid and LR+CV methods, the CFD method cannot reuse paths for different gradient estimates since the simulated paths have only one particular parameter perturbed. This problem with finite difference methods grows in significance as the dimension of θ grows.

5 Conclusions

We have provided a new class of methods for the estimation of parametric sensitivities. These hybrid methods include a pathwise estimate but also a correction term, ensuring that the bias is either mitigated (in the case of the RPD hybrid method) or zero. In particular, the GS hybrid method is, along with the LR and PPA methods, only the third unbiased method so far developed in the current setting for the estimation of derivatives of the form $\frac{\partial}{\partial \theta_i}\mathbb{E}[f(X_{\theta}(t))]$.

For computing sensitivities of the form $\frac{\partial}{\partial \theta}\mathbb{E}[f(X_{\theta}(t))]$ at some fixed time t , two methods were highlighted. The GS hybrid method is unbiased, and can be significantly more efficient than existing unbiased methods. At the cost of a small, controllable bias, the RPD hybrid method, which utilizes the RPD method of [30] for the pathwise estimate, can often increase efficiency further, particularly at large times when the system may be nearing stationarity. A useful avenue of future work will be to study these and other existing sensitivity methods on a wider range of networks and parameter values to better describe which method might be most efficient for a given model of interest.

Acknowledgments. Anderson and Wolf were both supported by NSF grant DMS-1318832. Anderson was also supported under Army Research Office grant W911NF-14-1-0401. We thank James Rawlings for

suggesting the study of Michaelis–Menten kinetics.

A Proof of Theorem 1

We restate Theorem 1.

Theorem 1. *Suppose that the process Z_θ satisfies the stochastic equation (1) with λ_k satisfying Conditions 3 and 4 on a neighborhood Θ of θ . Suppose that the function F satisfies Condition 2 on Θ . For some $0 \leq a \leq b < \infty$, let $L_Z(\theta) = \int_a^b F(\theta, Z_\theta(s)) ds$. Then $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_Z(\theta)] = \mathbb{E}\left[\frac{\partial}{\partial \theta_i} L_Z(\theta)\right]$, for all $i \in \{1, \dots, R\}$.*

The proof of Theorem 1 is similar to that of Theorem 5.1 in [15]. The main difference is in the proof of the continuity of the function L , which is our Lemma 2 below. As in Section 2.2.3, for convenience throughout this appendix we take $R = 1$ (so that θ is 1-dimensional).

We first need some preliminary results. Let $N(\theta, t)$ be the number of jumps of Z_θ through time t .

Lemma 1. *For any fixed and finite $t, q \in [1, \infty)$, and $c \in [1, \infty)$, we have*

$$\mathbb{E}\left[\sup_{\theta \in \Theta} N(\theta, t)^q\right] < \infty, \quad \mathbb{E}\left[\sup_{\theta \in \Theta} \sup_{s \in [0, t]} \|Z_\theta(s)\|^q\right] < \infty \quad \text{and} \quad \mathbb{E}\left[\sup_{\theta \in \Theta} c^{N(\theta, t)}\right] < \infty.$$

Proof. Note that by Condition 4, $N(\theta, t)$ is stochastically bounded, uniformly in θ , by a Poisson random variable \hat{N} with parameter $\tilde{\Gamma} = tK\Gamma_M$. This proves the first bound immediately. To see the second result, note that $\sup_{s \in [0, t]} \|Z_\theta(s)\| \leq \|Z_\theta(0)\| + N(\theta, t) \max_k |\mathbf{1} \cdot \zeta_k|$ and use the first result. To prove the final bound, use that $\mathbb{E}[\sup_{\theta \in \Theta} c^{N(\theta, t)}] \leq \mathbb{E}[c^{\hat{N}}]$, and that

$$\mathbb{E}[c^{\hat{N}}] = \sum_{m=0}^{\infty} c^m \mathbb{P}(\hat{N} = m) = \sum_{m=0}^{\infty} c^m \frac{\tilde{\Gamma}^m}{m!} e^{-\tilde{\Gamma}} = e^{-\tilde{\Gamma}} \sum_{m=0}^{\infty} \frac{(c\tilde{\Gamma})^m}{m!} = e^{-\tilde{\Gamma}} e^{c\tilde{\Gamma}} < \infty.$$

□

Lemma 2. *For any $\theta \in \Theta$ and for $h > 0$ such that $(\theta - h, \theta + h) \subset \Theta$, with probability $1 - O(h^2)$ we have that $L_Z(\theta)$ is continuous and piecewise differentiable on $(\theta - h, \theta + h)$.*

Proof. There are two parts to the proof. First, we show that if on the interval $(\theta - h, \theta + h)$ no more than one change occurs to the embedded chain \hat{Z}_ℓ on the interval $[a, b]$, then $L_Z(\theta)$ is continuous on that interval. Second, we require that the probability of two or more such changes is $O(h^2)$. The proof of the second claim follows as in the second part of Appendix 5.B in [15], p. 120, so we do not include it here.

We prove the first claim. Suppose that there is at most one change to the embedded chain in the time interval $[a, b]$ on $(\theta - h, \theta + h)$. Then one of the following cases occurs:

- (i) there is no change to the embedded chain,
- (ii) two (or more) jumps switch order through time b , causing a change in the embedded chain of Z_θ , or
- (iii) some jump enters or exits the interval $[a, b]$, changing the number states appearing in the integral L_Z .

We have crucially used the non-interruptive Condition 3 here, and the fact that Z_θ satisfies the random time change representation (1), to exclude any other possibilities, including interruptions. What we must show is that L_Z is continuous in each case. Recall from (16) that

$$L_Z(\theta) = \sum_{\ell=0}^{N(\theta, b)} F(\theta, \hat{Z}_\ell(\theta)) [T_{\ell+1}^\theta \wedge b - T_\ell^\theta \vee a]^+ \quad (36)$$

and that F is continuous in θ by assumption. By work in Section 2.2.3, the jump times T_ℓ^θ are continuous except possibly at values of θ at which the embedded chain of Z_θ changes. Thus it is clear that L_Z is continuous in case (i).

Now suppose that (ii) occurs at some point $\theta^* \in (\theta - h, \theta + h)$. Then two reactions k and m occur at the same time. (The case when three or more reactions occur simultaneously is essentially the same.) Further suppose these reactions occur as the ℓ^{th} and $(\ell + 1)^{\text{st}}$ jumps. Then at θ^* , there is a discontinuity in $\hat{Z}_\ell(\theta)$: from one side the limit is $\hat{Z}_{\ell-1}(\theta) + \zeta_k$ and from the other it is $\hat{Z}_{\ell-1}(\theta) + \zeta_m$. However, by the non-interruptive Condition, the two reactions may occur in either order, and the net result of the two reactions is the same regardless: $\zeta_k + \zeta_m$ is added to the system. That is, $\hat{X}_{\ell+1}(\theta) \equiv Z_{\ell-1}(\theta) + \zeta_k + \zeta_m$ on the whole interval, and furthermore, this crossover of jumps affects no other states of the embedded chain.

Then in the summation (36), any given term changes continuously except possibly the ℓ^{th} term,

$$F(\theta, \hat{Z}_\ell(\theta)) [T_{\ell+1}^\theta \wedge b - T_\ell^\theta \vee a]^+. \quad (37)$$

But at θ^* , we have $T_{\ell+1}^{\theta^*} = T_\ell^{\theta^*}$. That is, neither reaction is postponed because the intensities of both are strictly positive. Therefore, the term (37) is zero at the point of discontinuity, and $L_Z(\theta)$ is continuous at θ^* as needed.

Suppose instead that at θ^* case (iii) occurs. Since an additional jump time appears in the interval $[a, b]$ at θ^* , an additional term may show up in the summation (36). However, this new jump time T_ℓ^θ must be equal to either a or b . Then $[T_{\ell+1}^\theta \wedge b - T_\ell^\theta \vee a]^+$ is zero, and L_Z is again continuous at θ^* .

Finally, L_Z is piecewise differentiable in each case. Indeed, by the derivations in Section 2.2.3, L_Z is differentiable except possibly at values of θ at which the embedded chain changes, and by assumption there is at most one such value. \square

We now prove two useful bounds before finally giving the proof of Theorem 1. For the remainder, we assume for convenience that Γ_M, Γ_m , and Γ' are at least 1.

Lemma 3. *For each ℓ from 0 to $N(\theta, b)$ we have*

$$M_\ell := \max_k \max_{j \leq \ell} \left| \frac{\partial}{\partial \theta} S_k^\theta(T_j^\theta) \right| \leq \Gamma' b (2\Gamma_M \Gamma_m)^\ell,$$

where Γ_M, Γ_m , and Γ' are as in Condition 4.

Proof. Consider (19) and (20) and recall that for each k we have $\frac{\partial}{\partial \theta} S_k^\theta(T_0^\theta) = 0$. Then

$$\left| \frac{\partial}{\partial \theta} \Delta_0^\theta \right| = \left| \frac{\Delta_0^\theta}{\lambda_{k_\ell}(\theta, \hat{Z}_\theta(0))} \frac{\partial}{\partial \theta} \lambda_{k_0}(\theta, \hat{Z}_\theta(0)) \right| \leq \Delta_0^\theta \Gamma' \Gamma_m.$$

Then for any k , we have

$$\frac{\partial}{\partial \theta} S_k^\theta(T_1^\theta) = \Delta_0^\theta \frac{\partial}{\partial \theta} \lambda_k(\theta, \hat{Z}_\theta(0)) + \lambda_k(\theta, \hat{Z}_\theta(0)) \frac{\partial}{\partial \theta} \Delta_0^\theta,$$

so that

$$M_1 = \max_k \left| \frac{\partial}{\partial \theta} S_k^\theta(T_1^\theta) \right| \leq \Delta_0^\theta \Gamma' + \Gamma_M \Delta_0^\theta \Gamma' \Gamma_m \leq 2\Gamma' \Gamma_m \Gamma_M \Delta_0^\theta.$$

Similarly, for a given ℓ we have

$$\begin{aligned} \left| \frac{\partial}{\partial \theta} \Delta_\ell^\theta \right| &\leq \left| \frac{\Delta_\ell^\theta}{\lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta))} \frac{\partial}{\partial \theta} \lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta)) \right| + \left| \lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta))^{-1} \frac{\partial}{\partial \theta} S_{k_\ell}^\theta(T_\ell^\theta) \right| \\ &\leq \Delta_\ell^\theta \Gamma' \Gamma_m + \Gamma_m M_{\ell-1}. \end{aligned}$$

Therefore, using that

$$\frac{\partial}{\partial \theta} S_k^\theta(T_\ell^\theta) = \frac{\partial}{\partial \theta} S_k^\theta(T_{\ell-1}^\theta) + \Delta_{\ell-1}^\theta \frac{\partial}{\partial \theta} \lambda_k(\theta, \hat{Z}_{\ell-1}(\theta)) + \lambda_k(\theta, \hat{Z}_{\ell-1}(\theta)) \frac{\partial}{\partial \theta} \Delta_{\ell-1}^\theta$$

and noticing that the M_ℓ are nondecreasing, we see that

$$\begin{aligned}
M_\ell &\leq M_{\ell-1} + \Gamma' \Delta_{\ell-1}^\theta + \Gamma_M \left| \frac{\partial}{\partial \theta} \Delta_{\ell-1}^\theta \right| \\
&\leq M_{\ell-1} + \Gamma' \Delta_{\ell-1}^\theta + \Gamma_M (\Delta_{\ell-1}^\theta \Gamma' \Gamma_m + \Gamma_m M_{\ell-2}) \\
&\leq M_{\ell-1} + \Gamma' \Delta_{\ell-1}^\theta + \Gamma_M (\Delta_{\ell-1}^\theta \Gamma' \Gamma_m + \Gamma_m M_{\ell-1}) \\
&\leq 2\Gamma_M \Gamma_m M_{\ell-1} + 2\Gamma' \Gamma_M \Gamma_m \Delta_{\ell-1}^\theta.
\end{aligned}$$

Iterating this inequality, we see that

$$M_\ell \leq (2\Gamma_M \Gamma_m)^{\ell-1} 2\Gamma' \Gamma_M \Gamma_m \sum_{j=0}^{\ell-1} \Delta_j^\theta \leq \Gamma' b (2\Gamma_M \Gamma_m)^\ell. \quad \square$$

Corollary 1. *For each ℓ from 0 to $N(\theta, b)$ we have*

$$\left| \frac{\partial}{\partial \theta} \Delta_\ell^\theta \right| \leq 2\Gamma' b \Gamma_m (2\Gamma_M \Gamma_m)^\ell,$$

where Γ_M, Γ_m , and Γ' are as in Condition 4.

Proof. By (19), the two final assumptions on Z_θ from Appendix A, and Lemma 3, we have that

$$\begin{aligned}
\left| \frac{\partial}{\partial \theta} \Delta_\ell^\theta \right| &\leq \left| \frac{\Delta_\ell^\theta}{\lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta))} \frac{\partial}{\partial \theta} \lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta)) \right| + \left| \lambda_{k_\ell}(\theta, \hat{Z}_\ell(\theta))^{-1} \frac{\partial}{\partial \theta} S_{k_\ell}^\theta(T_\ell^\theta) \right| \\
&\leq b\Gamma_m \Gamma' + \Gamma_m \left| \frac{\partial}{\partial \theta} S_{k_\ell}^\theta(T_\ell^\theta) \right| \\
&\leq b\Gamma_m \Gamma' + \Gamma_m \Gamma' b (2\Gamma_M \Gamma_m)^\ell \\
&\leq 2\Gamma' b \Gamma_m (2\Gamma_M \Gamma_m)^\ell.
\end{aligned}$$

□

We finally turn to the proof of Theorem 1. As noted previously, the proof of the theorem now follows similarly to the proof of Theorem 5.1 in [15].

Proof of Theorem 1. Let \tilde{h} be the infimum over h for which two or more changes occur to the embedded chain of Z_θ through $(\theta - h, \theta + h)$ on the time interval $[a, b]$. That is, \tilde{h} is the *second* place at which a change in the embedded chain occurs. Note that $\tilde{h} > 0$ is positive with probability 1. Without loss of generality, $(\theta - \tilde{h}, \theta + \tilde{h}) \subset \Theta$. We must prove the middle equality in

$$\frac{d}{d\theta} \mathbb{E}[L_Z(\theta)] = \lim_{h \rightarrow 0} \mathbb{E}[h^{-1} [L_Z(\theta + h) - L_Z(\theta)]] = \mathbb{E} \left[\lim_{h \rightarrow 0} h^{-1} [L_Z(\theta + h) - L_Z(\theta)] \right] = \mathbb{E} \left[\frac{d}{d\theta} L_Z(\theta) \right].$$

We write

$$\begin{aligned}
&\mathbb{E}[h^{-1} [L_Z(\theta + h) - L_Z(\theta)]] \\
&= \mathbb{E}[h^{-1} [L_Z(\theta + h) - L_Z(\theta)] \mathbf{1}(h < \tilde{h})] + \mathbb{E}[h^{-1} [L_Z(\theta + h) - L_Z(\theta)] \mathbf{1}(h \geq \tilde{h})].
\end{aligned} \tag{38}$$

Consider the first term. By Lemma 2, and since by the definition of \tilde{h} at most one change occurs to the embedded chain for $h < \tilde{h}$, we have that L_Z is continuous and piecewise differentiable on $(\theta - \tilde{h}, \theta + \tilde{h})$. By a generalized mean value theorem (e.g. [9]),

$$|h^{-1} [L_Z(\theta + h) - L_Z(\theta)] \mathbf{1}(h < \tilde{h})| \leq \sup_{\theta \in \Theta} \left| \frac{d}{d\theta} L_Z(\theta) \right|,$$

where the supremum is over those points where the derivative exists. We will show that this supremum has finite expectation; therefore, since as $h \rightarrow 0$,

$$h^{-1}[L_Z(\theta + h) - L_Z(\theta)]\mathbf{1}(h < \tilde{h}) \xrightarrow{a.s.} \frac{d}{d\theta}L_Z(\theta)$$

we will have by the dominated convergence theorem that $\mathbb{E}[h^{-1}[L_Z(\theta + h) - L_Z(\theta)]\mathbf{1}(h < \tilde{h})] \rightarrow \mathbb{E}[\frac{d}{d\theta}L_Z(\theta)]$. We will also show that the second term in (38) goes to zero as $h \rightarrow 0$, which proves the theorem.

Write $N := N(\theta, b)$ and recall that

$$\begin{aligned} \left| \frac{d}{d\theta}L_Z(\theta) \right| &= \left| \sum_{\ell=0}^N [T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \left(\frac{\partial}{\partial\theta}F(\theta, \hat{Z}_\ell(\theta)) \right) + F(\theta, \hat{Z}_\ell(\theta)) \frac{\partial}{\partial\theta}[T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \right| \\ &\leq \left| \sum_{\ell=0}^N [T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \left(\frac{\partial}{\partial\theta}F(\theta, \hat{Z}_\ell(\theta)) \right) \right| + \left| \sum_{\ell=0}^N F(\theta, \hat{Z}_\ell(\theta)) \frac{\partial}{\partial\theta}[T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \right|. \end{aligned}$$

We now consider these two terms separately. By Condition 2 on F ,

$$\begin{aligned} \left| \sum_{\ell=0}^N [T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \left(\frac{\partial}{\partial\theta}F(\theta, \hat{Z}_\ell(\theta)) \right) \right| &\leq \sum_{\ell=0}^N [T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \left| \frac{\partial}{\partial\theta}F(\theta, \hat{Z}_\ell(\theta)) \right| \\ &\leq C_2 \sum_{\ell=0}^N [T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ (1 + \|\hat{Z}_\ell^\theta\|^{c_2}) \\ &\leq C_2(b-a)(1 + \max_{\ell \leq N} \|\hat{Z}_\ell^\theta\|^{c_2}) \\ &\leq C_2(b-a)(1 + \sup_{\theta \in \Theta} \sup_{s \in [0, b]} \|Z_\theta(s)\|^{c_2}). \end{aligned}$$

Now, from (21) and our work in Lemma 3 we have for any ℓ that

$$\left| \frac{\partial}{\partial\theta}[T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \right| \leq \sum_{j=0}^N \left| \frac{\partial}{\partial\theta}\Delta_j \right|.$$

Therefore, for the second term,

$$\begin{aligned} \left| \sum_{\ell=0}^N F(\theta, \hat{Z}_\ell(\theta)) \frac{\partial}{\partial\theta}[T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \right| &\leq C_1 \sum_{\ell=0}^N (1 + \|\hat{Z}_\ell^\theta\|^{c_1}) \left| \frac{\partial}{\partial\theta}[T_{\ell+1}(\theta) \wedge b - T_\ell(\theta) \vee a]^+ \right| \\ &\leq C_1(1 + \max_{\ell \leq N} \|\hat{Z}_\ell^\theta\|^{c_1}) \sum_{\ell=0}^N \sum_{j=0}^N \left| \frac{\partial}{\partial\theta}\Delta_j \right| \\ &\leq C_1(1 + \max_{\ell \leq N} \|\hat{Z}_\ell^\theta\|^{c_1}) \sum_{\ell=0}^N \sum_{j=0}^N 2\Gamma' T \Gamma_m^2 (2\Gamma_M \Gamma_m)^j \\ &\leq C_1(1 + \sup_{\theta \in \Theta} \sup_{s \in [0, b]} \|Z_\theta(s)\|^{c_1}) N^2 2\Gamma' T \Gamma_m (2\Gamma_M \Gamma_m)^N. \end{aligned}$$

By Lemma 1 and repeated applications of the Cauchy-Schwarz inequality, we see that both of the bounds we have computed are bounded uniformly in θ on Θ by a quantity of finite expectation as needed.

Finally, we must show that $\mathbb{E}[h^{-1}[L_Z(\theta + h) - L_Z(\theta)]\mathbf{1}(h \geq \tilde{h})]$ goes to zero as $h \rightarrow 0$. By using the Cauchy-Schwarz inequality, we see that

$$\mathbb{E} \left[h^{-1}[L_Z(\theta + h) - L_Z(\theta)]\mathbf{1}(h \geq \tilde{h}) \right]^2 \leq h^{-2} \mathbb{E} [L_Z(\theta + h) - L_Z(\theta)]^2 P(h \geq \tilde{h}).$$

Since by Lemma 2 we have $P(h \geq \tilde{h}) = O(h^2)$, and since $[L_Z(\theta + h) - L_Z(\theta)] \xrightarrow{a.s.} 0$, we are done by the dominated convergence theorem if we can show that $[L_Z(\theta + h) - L_Z(\theta)]^2$ is bounded by an integrable function. By Condition 2 on F , for any $\theta \in \Theta$,

$$\begin{aligned} [L_Z(\theta)]^2 &= \left(\int_a^b F(\theta, Z_\theta(s)) ds \right)^2 \leq (b-a) \int_a^b (F(\theta, Z_\theta(s)))^2 ds \\ &\leq (b-a) \int_a^b C_1^2 (1 + \|Z_\theta(s)\|^{c_1})^2 ds \\ &\leq C_1^2 (b-a)^2 (2 + 2 \sup_{\theta \in \Theta} \sup_{s \in [0, b]} \|Z_\theta(s)\|^{2c_1}), \end{aligned} \tag{39}$$

where the final line follows because $(a+b)^2 \leq 2a^2 + 2b^2$. This bound has finite expectation by Lemma 1, and is also uniform, so that it holds for $|L_Z(\theta + h)|$ as well. Then as needed,

$$|L_Z(\theta + h) - L_Z(\theta)|^2 \leq 2[L_Z(\theta + h)]^2 + 2[L_Z(\theta)]^2 \leq 4 \sup_{\theta \in \Theta} [L_Z(\theta)]^2,$$

which has finite expectation by taking the supremum of (39). \square

References

- [1] David F. Anderson, *A modified next reaction method for simulating chemical systems with time dependent propensities and delays*, J. Chem. Phys. **127** (2007), no. 21, 214107.
- [2] ———, *An efficient finite difference method for parameter sensitivities of continuous time Markov chains*, SIAM: Journal on Numerical Analysis **50** (2012), 2237–2258.
- [3] David F. Anderson, Bard Ermentrout, and Peter J. Thomas, *Stochastic representations of ion channel kinetics and exact stochastic simulation of neuronal dynamics*, accepted for publication to Journal for Computational Neuroscience, 2014.
- [4] David F. Anderson and Desmond J. Higham, *Multilevel monte carlo for continuous time markov chains, with applications in biochemical kinetics*, Multiscale Modeling & Simulation **10** (2012), no. 1, 146–179.
- [5] David F. Anderson and Masanori Koyama, *An asymptotic relationship between coupling methods for stochastically modeled population processes*, accepted to IMA Journal of Numerical Analysis, 2014.
- [6] David F. Anderson and Thomas G. Kurtz, *Stochastic analysis of biochemical systems*, Springer, MBI series, to appear.
- [7] ———, *Continuous time Markov chain models for chemical reaction networks*, Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology (H. Koeppl et al., ed.), Springer, 2011, pp. 3–42.
- [8] Soren Asmussen and Peter W. Glynn, *Stochastic simulation: Algorithms and analysis*, Springer, 2007.
- [9] Jean Alexandre Dieudonné, Jean Dieudonné, France Mathematician, and Jean Dieudonné, *Foundations of modern analysis*, vol. 286, Academic press New York, 1960.
- [10] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain, *Stochastic gene expression in a single cell*, Science **297** (2002), no. 5584, 1183–1186.
- [11] Stewart N. Ethier and Thomas G. Kurtz, *Markov processes: Characterization and convergence*, 2 ed., John Wiley & Sons, New York, 2005.
- [12] M.A. Gibson and J. Bruck, *Efficient exact stochastic simulation of chemical systems with many species and many channels*, J. Phys. Chem. A **105** (2000), 1876–1889.

- [13] Mike B. Giles, *Multilevel Monte Carlo path simulation*, Operations Research **56** (2008), 607–617.
- [14] Daniel T. Gillespie, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*, J. Comput. Phys. **22** (1976), 403–434.
- [15] Paul Glasserman, *Gradient estimation via perturbation analysis*, Kluwer Academic Publishers, 1991.
- [16] Peter W Glynn, *Likelihood ratio gradient estimation for stochastic systems*, Communications of the ACM **33** (1990), no. 10, 75–84.
- [17] Wei-Bo Gong and Yu-Chi Ho, *Smoothed (conditional) perturbation analysis of discrete event dynamical systems*, Automatic Control, IEEE Transactions on **32** (1987), no. 10, 858–866.
- [18] Ankit Gupta and Mustafa Khammash, *Unbiased estimation of parameter sensitivities for stochastic chemical reaction networks*, SIAM Journal on Scientific Computing **35** (2013), no. 6, A2598–A2620.
- [19] Ankit Gupta and Mustafa Khammash, *An efficient and unbiased method for sensitivity analysis of stochastic reaction networks*, Royal Society Interface **11** (2014), no. 101, 20140979.
- [20] ———, *Sensitivity analysis for stochastic chemical reaction networks with multiple time-scales*, Electronic Journal of Probability **19** (2014), no. 59, 1–53.
- [21] Thomas G. Kurtz, *Strong approximation theorems for density dependent Markov chains*, Stoch. Proc. Appl. **6** (1978), 223–240.
- [22] ———, *Representations of Markov processes as multiparameter time changes*, Ann. Prob. **8** (1980), no. 4, 682–715.
- [23] ———, *Representation and approximation of counting processes*, Advances in filtering and optimal stochastic control, Lecture Notes in Control and Information Sciences, vol. 42, Springer, Berlin, 1982, pp. 177–191.
- [24] Brian Munsky and Mustafa Khammash, *The finite state projection algorithm for the solution of the chemical master equation*, The Journal of chemical physics **124** (2006), no. 4, 044104.
- [25] Johan Paullson, *Summing up the noise in gene networks*, Nature **427** (2004), 415–418.
- [26] Sergey Plyasunov and Adam P. Arkin, *Efficient stochastic sensitivity analysis of discrete event systems*, J. Comp. Phys. **221** (2007), 724 – 738.
- [27] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi, *Stochastic mRNA synthesis in mammalian cells*, PLoS biology **4** (2006), no. 10, e309.
- [28] Muruhan Rathinam, Patrick W. Sheppard, and Mustafa Khammash, *Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks*, Journal of Chemical Physics **132** (2010), 034103.
- [29] Kevin R. Sanft, Daniel T. Gillespie, and Linda R. Petzold, *Legitimacy of the stochastic Michaelis Menten approximation*, Systems Biology, IET **5** (2011), no. 1, 58–69.
- [30] Patrick W. Sheppard, Muruhan Rathinam, and Mustafa Khammash, *A pathwise derivative approach to the computation of parameter sensitivities in discrete stochastic chemical systems.*, The Journal of chemical physics **136** (2012), no. 3, 034115.
- [31] Rishi Srivastava, David F Anderson, and James B Rawlings, *Comparison of finite difference based methods to obtain sensitivities of stochastic chemical kinetic models*, The Journal of chemical physics **138** (2013), no. 7, 074110.
- [32] Darren J. Wilkinson, *Stochastic modelling for systems biology*, second ed., Chapman and Hall/CRC Press, 2011.

- [33] Elizabeth Skubak Wolf and David F. Anderson, *A finite difference method for estimating second order parameter sensitivities of discrete stochastic chemical reaction networks*, J. Chem. Phys. **137** (2012), no. 22, 224112.